ITERATIVE TIMESTAMPED AUDIO DESCRIBER

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world audio applications require precise, queryable representations grounded in physical time. Yet modern Large Audio-Language Models (LALMs) increasingly favor conversational generation, producing free-form descriptions that conflate acoustic perception with higher-level reasoning and lack precise temporal grounding. We introduce the Iterative Timestamped Audio Describer (`ITAD`), a foundation model for audio perception that constructs granular, speaker-attributed acoustic timelines directly from in-the-wild audio. To avoid prohibitively expensive manual annotation, we programmatically fuse complementary predictions from specialized weak teachers through time-registered IoU matching. We model the resulting multi-granularity schema of timestamped events, speaker turns, and fine-grained speech attributes using an Instruction-Input-Output (IIO) completion format, enabling both autonomous decoding and the injection of external priors through a dynamic control surface. We further extend `ITAD` with expressive speech attributes to obtain `ITAD-S`, using parameter-efficient specialists to elicit latent paralinguistic capabilities from the pre-trained backbone and distill them into a unified model. We validate the resulting representation in two complementary settings. Paired with frozen text-only LLMs, it yields strong spoken audio understanding across diverse reasoning backends, demonstrating the utility of decoupling fine-grained acoustic perception from downstream reasoning. Used as paired supervision for text-to-audio generation, it supports semantically and temporally grounded reconstruction, together with explicit zero-shot control over event timing, speaker attributes, and word-level emphasis. Project page: `https://spbanonymo.github.io/itad-audio-describer/`

1 INTRODUCTION

Human auditory perception separates overlapping acoustic signals before higher-level reasoning (Bregman, 1990), grounding abstract semantics in physical perception (Harnad, 1990; Bender & Koller, 2020). In contrast, modern Large Audio Language Models (LALMs) emphasize conversational generation (Tang et al., 2024; Chu et al., 2024; Xu et al., 2025), producing free-form text with limited structural grounding. Yet applications such as semantic search, precise audio editing, and automated archiving require queryable, temporally grounded representations of events, speakers, and paralinguistic cues.

We propose the **Iterative Timestamped Audio Describer** (`ITAD`), a foundation model that constructs granular, speaker-attributed acoustic timelines directly from in-the-wild audio. To avoid manually annotating such structured timelines at scale, we introduce a multi-view fusion pipeline that temporally aligns complementary predictions from off-the-shelf sound event detectors, diarizers, and captioners over a large real-world corpus, producing high-fidelity structural training targets. To model these targets while retaining modular control, we introduce an **Instruction-Input-Output (IIO)** completion format that decomposes the schema into discrete tasks with optional teacher-forced inputs. IIO supports both autonomous scene decoding and external priors such as high-quality transcripts, enabling iterative bootstrapping and controllable inference. We further extend `ITAD` with expressive speech attributes to obtain `ITAD-S`, formulating this expansion as latent capability elicitation. Motivated by probing evidence that these attributes are decodable from intermediate backbone representations, we train attention-only specialists on fewer than 10,000 labeled samples to expose them as structured outputs. These specialists annotate the corpus with expressive attributes, and the resulting cascade is distilled into the unified `ITAD-S` model.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

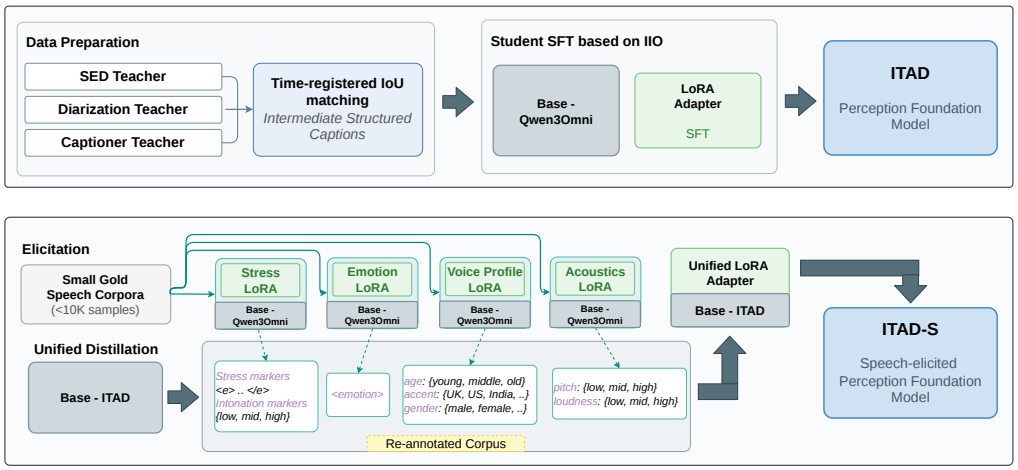

Figure 1: **Overview of the two-stage `ITAD` / `ITAD-S` framework. (1) Perceptual Foundation (`ITAD`):** Predictions from weak teachers (SED, Diarization, Captioner) are merged via time-registered IoU matching to generate intermediate structured captions, which train a base adapter via the proposed Instruction-Input-Output (IIO) format. **(2) Speech Elicitation (`ITAD-S`):** Individual LoRA specialists (Stress, Emotion, Voice Profile, Acoustics) are trained on small gold corpora to elicit paralinguistic attributes from the pre-trained backbone. This cascade re-annotates the corpus, which is subsequently distilled into our unified student model, `ITAD-S`.

The resulting structured representation enables two downstream paradigms. First, cascading `ITAD-S` with frozen text-only LLMs yields strong spoken understanding on MMSU (Wang et al., 2026a) and MSU-Bench (Wang et al., 2026c) across diverse reasoning backends, bridging perception and reasoning through a structured intermediate representation. Second, using `ITAD-S` to annotate in-the-wild audio produces a large-scale, paralinguistically enriched structured training corpus for text-to-audio generation. We show that this corpus enables explicit zero-shot control over event timing, speaker characteristics, and word-level emphasis.

In this work, our contributions are: (1) a scalable supervision framework that constructs granular, temporally grounded training targets by fusing complementary weak teachers, and extends them with paralinguistic attributes elicited by lightweight specialists trained on small labeled datasets (2) Instruction-Input-Output (IIO), a modular structured-completion interface that supports both autonomous decoding and externally conditioned inference, naturally enabling iterative annotation and bootstrapping; and (3) `ITAD` and its paralinguistic extension `ITAD-S`, which instantiate this framework and demonstrate strong fine-grained audio understanding together with temporally and paralinguistically controllable audio generation.

2 Related Work

Audio-Language Models and Grounded Perception. Recent Large Audio-Language Models (LALMs) (Tang et al., 2024; Chu et al., 2024; Kong et al., 2024; Zhang et al., 2025) demonstrate strong instruction-following, yet benchmarks reveal a persistent gap between high-level reasoning and fine-grained acoustic perception (Wang et al., 2026a; Sakshi et al., 2025). Existing models remain susceptible to temporal hallucinations and reliance on linguistic priors (Kuan & Lee, 2025; Cheng et al., 2026), a limitation compounded by captioning datasets that summarize audio into short global descriptions (Kim et al., 2019; Drossos et al., 2020). While prior timestamped captioners such as TAC (Kumar et al., 2026) introduced explicit event boundaries, they primarily target event-level temporal grounding and are not designed to represent the richer structure of natural conversational audio, including speaker-attributed turns and overlapping events. In contrast, `ITAD` uses multi-teacher temporal fusion to construct explicit, speaker-attributed acoustic timelines from in-the-wild audio, and represents them through IIO as a multi-granularity perceptual interface that supports both autonomous decoding and externally conditioned inference.

Spoken Paralinguistics and Latent Elicitation. Expressive speech attributes such as stress, intonation, emotion, and speaker characteristics are central to spoken understanding (Nguyen et al., 2023; Jin et al., 2024). Existing SpeechLLMs often model these attributes as isolated categorical tasks (Wang et al., 2024; Geng et al., 2025) or rely on external signal-processing pipelines for acoustic features (Yosha et al., 2025; Wang et al., 2026b). Meanwhile, work on latent capability elicitation suggests that pretrained models can encode capabilities that become accessible through limited parameter adaptation (Aghajanyan et al., 2021; Donoway et al., 2026). Building on this observation, ITAD−S uses lightweight specialists to expose paralinguistic information that is decodable from the pretrained audio backbone, and distills their predictions into a unified, multi-granularity acoustic schema.

3 METHODOLOGY

Our framework constructs a queryable, temporally grounded acoustic representation in two stages. Stage 1 builds ITAD, a granular, speaker-attributed physical timeline bootstrapped from complementary weak teachers over in-the-wild audio. Stage 2 extends this timeline into ITAD−S by eliciting latent paralinguistic capabilities from the pre-trained backbone using lightweight specialists, then distilling their annotations into a unified model.

3.1 STAGE 1: PERCEPTUAL FOUNDATION (ITAD)

Dense temporal annotation at the scale required to train a foundation model is prohibitively expensive. We therefore construct ITAD's training targets through multi-view weak supervision and learn them using an Instruction-Input-Output (IIO) completion paradigm.

Multi-view Perceptual Fusion We programmatically construct ITAD's target timelines by fusing predictions from specialized weak teachers, including audio captioning, speaker diarization, and sound event detection. While each teacher provides complementary task-specific predictions under a distinct schema, all predictions share physical time as a common coordinate. We exploit this common axis to align their disparate temporal boundaries and attributes (Figure 2).

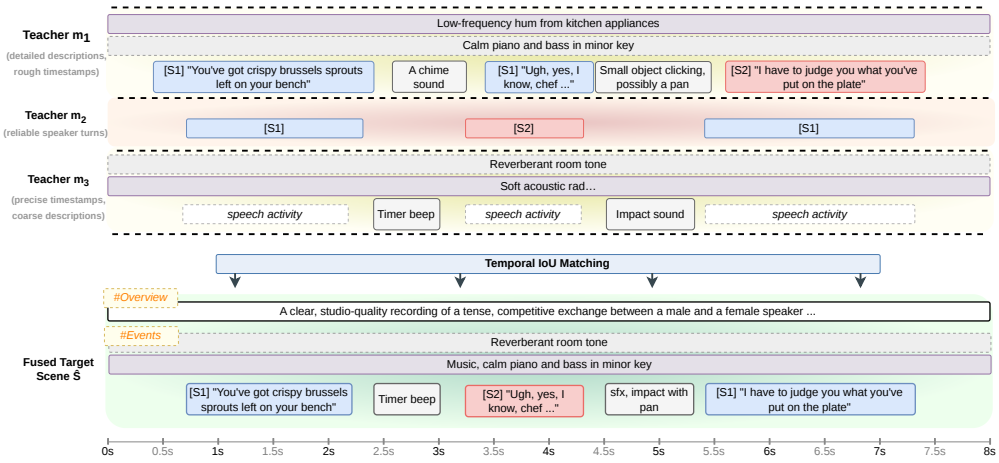

Figure 2: **Programmatic data compilation via time-registered IoU matching.** Predictions from three independent weak teachers (m_1: audio captioner; m_2: speaker diarizer; m_3: sound event detector) are aligned on a shared time axis using temporal IoU with overlap gate (Eq. 1,2). Their aligned predictions form the fused target timeline $\hat{S}$ (Eq. 3), containing concurrent acoustic events with speaker attribution.

Let the latent scene S comprise acoustic events over the attribute schema $\mathcal{A} = \{\text{event type}, \text{acoustic description}, \text{temporal boundaries}, \text{speaker identity}, \text{transcript}\}$. Each of K specialized teachers m_k produces a partial view $O_k = o$, with a temporal support $\tau(o) = [o_s, o_e]$.

To fuse these views, we define an authority map auth : $\mathcal{A} \to \{1, \ldots, K\}$ that maps each attribute to its most reliable teacher. Furthermore, each acoustic event class c is assigned a temporal owner $\kappa(c)$ that dictates its master temporal segmentation. Temporal agreement between two predictions is measured using their temporal Intersection over Union (IoU):

$$\mathrm{IoU}(o, o') = \frac{|\tau(o) \cap \tau(o')|}{|\tau(o) \cup \tau(o')|} \tag{1}$$

For each event class c, let $E_c = (e_1, \ldots, e_{n_c})$ be its reference events drawn from its temporal owner $\kappa(c)$. Against the time-ordered items of view O_k, we compute a one-to-one partial alignment $\pi_k^{(c)} : E_c \to O_k \cup \emptyset$ that maximizes total temporal overlap over matched references:

$$\pi_k^{(c)} = \arg\max_{\pi} \sum_{\substack{e \in E_c \\ \pi(e) \neq \emptyset}} \mathrm{IoU}(e, \pi(e)) \quad \text{s.t.} \quad \mathrm{IoU}(e, \pi(e)) \geq \tau_k \;\; \forall\, e : \pi(e) \neq \emptyset \tag{2}$$

where τ_k is minimum temporal overlap threshold. We then construct the fused target timeline $\hat{S}$ by taking each event's temporal span from its temporal owner and each attribute from its designated authority:

$$\hat{S} = \bigcup_{c} \left\{ e \;\middle|\; \forall \alpha \in \mathcal{A},\; \tau(e) = \tau(e^{\kappa(c)}),\; \alpha(e) = g_\alpha\left(\pi^{(c)}_{\mathrm{auth}(\alpha)}(e)\right) \right\} \tag{3}$$

where $e^{\kappa(c)}$ represents the class c events of the temporal owner, and g_α is a deterministic attribute selection function (for instance, longest string for open-text descriptions) that resolves cases where an annotator contributes multiple candidate values.

Instruction Input Output completion format Decoding a nested, multi-granularity timeline in a single pass imposes an all-or-nothing inference paradigm. In practice, however, downstream users may already possess high-quality partial signals such as an expert human transcript or output from a specialized diarization system. A standard end-to-end captioner cannot naturally incorporate these external priors to condition its remaining predictions, limiting its use as a modular component in larger audio pipelines.

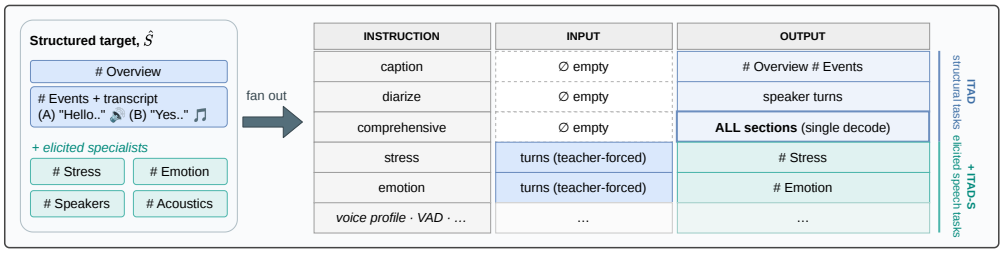

Figure 3: **The IIO completion format.** The fused target $\hat{S}$ is decomposed into (`INSTRUCTION`, `INPUT`, `OUTPUT`) completion tasks. The `INSTRUCTION` specifies the task, `INPUT` provides optional teacher-forced context, and `OUTPUT` specifies the target attributes to generate. Structural tasks train `ITAD` (`caption`, `diarize`, and a single-decode `comprehensive` task). Appending elicited speech tasks (`stress`, `emotion`, `voice profile`, `acoustics`) seamlessly extends this format to `ITAD-S` without changing the training procedure. At inference, the `comprehensive` task supports single-pass generation, while sequencing attribute-specific tasks with preceding outputs yields a modular multi-pass cascade.

We address this limitation with an *Instruction-Input-Output* (IIO) completion format, which fans out the target schema into completion tasks over the same audio (Figure 3). The `INSTRUCTION` specifies the task, the `INPUT` is either empty or contains teacher-forced context such as a transcript or speaker turns, and the `OUTPUT` is the remaining schema that we expect the model to generate. We train `ITAD` as a LoRA adapter ΔW_{base} on a frozen pre-trained backbone W_0, with adapted parameters $\theta_{\text{base}} = W_0 + \Delta W_{\text{base}}$, by minimizing the completion loss over the fanned-out tasks:

$$\mathcal{L}_{\text{base}} = -\sum_{t=1}^{|Y_{\text{base}}|} \log P\left(y_t^{\text{base}} \;\middle|\; Z_a,\; y_{<t}^{\text{base}},\; X_{\text{inst}},\; X_{\text{input}};\; W_0 + \Delta W_{\text{base}}\right) \tag{4}$$

where Z_a is the audio, X_{inst} the instruction, X_{input} the optional teacher-forced context, and Y_{base} the target completion.

Two properties make IIO central to our approach. Firstly, it is *extensible*: new attributes enter the schema as additional completion tasks without changing the training procedure. Stage 2 (Section 3.2) exploits this property by appending elicited speech attributes to the same task fan-out (Figure 3, + `ITAD-S` rows). Second, IIO exposes a *dynamic control surface* at inference. With an empty `INPUT`, the `comprehensive` task autonomously generates the full acoustic timeline in a single pass. Alternatively, attribute-specific tasks can consume preceding predictions or externally supplied priors through `INPUT`, enabling an iterative multi-pass cascade. For example, injecting a high-quality transcript from an external ASR system allows subsequent attributes to be predicted while holding the transcript fixed. This same mechanism enables iterative data bootstrapping as stronger external priors become available. We compare these inference modes in Section 4.1 and detail the per-task sampling mixture in Appendix D.

3.2 Stage 2: Speech-centric Elicitation (`ITAD-S`)

We next extend `ITAD`'s physical timeline with expressive speech attributes, including word stress, intonation, emotion, and speaker characteristics. Rather than learning these capabilities from scratch at scale, we elicit paralinguistic information already encoded within the pre-trained backbone using lightweight specialists, then distill their annotations into the unified model, `ITAD-S`.

Parameter-efficient Specialist Elicitation Clean paralinguistic labels are scarce at the scale required for our in-the-wild corpus. We therefore formulate this expansion as *latent capability elicitation*. Pre-trained models can encode capabilities internally that are not reliably expressed in their outputs, but can be surfaced through limited parameter updates (Donoway et al., 2026). We hypothesize that the pre-trained LALM backbone similarly encodes expressive acoustic traits that are poorly exposed through its textual generation.

We test this hypothesis through layer-wise linear probing of the intermediate hidden states h_ℓ of the Qwen3Omni-captioner model. For a binary expressive attribute $c_i \in \{0, 1\}$, we fit an L_2 regularized logistic probe g_ℓ at every layer $\ell \in [0, L]$ by minimizing the cross-entropy:

$$\min_{w_\ell, b_\ell} -\sum_{i=1}^{|\mathcal{D}|} \Big[c_i \log \sigma(z_{\ell,i}) + (1 - c_i) \log \big(1 - \sigma(z_{\ell,i})\big) \Big] + \lambda \|w_\ell\|_2^2 \tag{5}$$

where $z_{\ell,i} = w_\ell^T h_{\ell,i} + b_\ell$ and $\sigma(z)$ is the sigmoid function. Our analysis reveals a gap between *latent decodability* and *generative accessibility*. Paralinguistic traits are strongly linearly decodable from the frozen backbone despite substantially weaker textual attribute generation 4.1. Specialist adaptation leaves early- and mid-layer decodability largely unchanged, with gains concentrated in later layers. Appendix C provides the layer-wise analysis.

Motivated by this latent perception, we perform sample-efficient acoustic elicitation. More concretely, instead of using large datasets to teach the model how to perceive prosody from scratch, we train attention-only Low-Rank Adaptation (LoRA) specialists on small gold-standard datasets (< 10K samples). Given the strong decodability of these attributes in the frozen backbone, these specialists primarily learn to expose the latent representations as explicit, structured output tags rather than learning the underlying acoustic features from scratch.

We train one LoRA specialist for each speech attribute $e \in \mathcal{E}$ (such as word stress, intonation, emotion, and voice profile) by adding a parameter efficient adapter $\Delta W_{\text{spec},e}$ to the frozen base W_0:

$$\mathcal{L}_{\text{spec},e} = -\sum_{t=1}^{|Y_e|} \log P\left(y_{t,e} \mid Z_a,\ y_{<t,e},\ X_{\text{inst},e},\ X_{\text{input},e};\ W_0 + \Delta W_{\text{spec},e}\right) \tag{6}$$

where $X_{\text{input},e}$ supplies a task-specific (teacher-forced) context, such as transcript or speaker list. The context allows each specialist to focus on predicting its target attribute Y_e rather than re-deriving the surrounding description. We detail the conditioning for specialists in Appendix B.

Once trained, we hot-load the specialists as co-existing LoRAs over the frozen Qwen3Omni-captioner to form our teacher cascade, `ITAD-E`. We deploy this over our in-the-wild corpus to augment the Stage 1 timelines with speech-centric, multi-attribute distillation targets.

Table 1: **Spoken and general audio understanding.** Models in **DtR ✓** rows emit a text schema as the description, which is read by a frozen text-only reasoner. Each score reflects the fidelity and coverage of audio perception in its generated description.

Model	DtR	MMSU			MMAR				MMAU				Avg.
		Perception	Reasoning	Overall	Speech	Sound	Music	Overall	Speech	Sound	Music	Overall	
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
StepAudio2	✗	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	77.4	66.2	66.0	58.8	45.2	60.1	67.0	70.0	71.3	69.4	65.2
Gemini-3.7 Flash	✓	68.7	84.4	76.3	82.7	51.5	56.8	69.5	80.5	66.1	67.1	71.2	72.3
Qwen3Omni-Capt.	✓	63.5	83.4	73.1	77.6	59.4	54.4	69.2	**84.1**	73.0	68.0	75.0	72.4
TAC	✓	63.8	84.8	74.0	83.7	**62.4**	**57.3**	**71.1**	82.0	69.4	69.8	73.7	72.9
ITAD-E	✓	**79.0**	**87.1**	**82.9**	**84.0**	56.4	52.9	70.3	80.5	71.8	68.3	73.5	**75.6**
ITAD-S	✓	77.8	86.4	81.9	79.6	53.3	51.0	68.0	82.3	73.9	70.7	**75.6**	75.2

Unified Student Distillation We distill ITAD-E into a single unified student, ITAD-S.

Starting from the Stage 1 ITAD weights θ_{base}, we learn a student LoRA ΔW_{stu}:

$$\mathcal{L}_{\text{dist}} = -\sum_{t=1}^{|Y_{\text{casc}}|} \log P_{\text{stu}}\left(y_t^{\text{casc}} \mid Z_a,\ y_{<t}^{\text{casc}},\ X_{\text{inst}},\ X_{\text{input}};\ \theta_{\text{base}} + \Delta W_{\text{stu}}\right) \tag{7}$$

where Y_{casc} denotes the enriched targets produced by ITAD-E. The paralinguistic attributes are incorporated as additional fanned-out IIO tasks mixed with the Stage 1 objectives, allowing ITAD-S to learn multi-granularity completion across the distillation target.

4 Experiments

A good describer should encode attributes relevant for both reasoning and generation in its text output. Here, we evaluate our trained captioners in these two settings: *audio understanding*, by cascading ITAD-S with a frozen text-only LLM which answers questions based on ITAD-S's generated description, and *audio generation*, by using its structured descriptions as paired supervision for text-to-audio generation.

4.1 Audio Understanding: validating schema fidelity

Under the *Describe-then-Reason (DtR)* protocol (Kumar et al., 2026), ITAD-S converts audio into a structured acoustic schema that a frozen text-only LLM uses to answer multiple-choice questions. This helps us assess whether the schema preserves the perceptual information required for downstream reasoning. We evaluate on MMSU (Wang et al., 2026a), the general audio benchmarks MMAR (Ma et al., 2026) and MMAU (Sakshi et al., 2025), and MSU-Bench (Wang et al., 2026c).

ITAD-S achieves an 81.9 overall score on MMSU, outperforming existing LALMs and nearing ITAD-E, the teacher cascade (82.9). The gain is concentrated in perception (Table 1): ITAD-S reaches 77.8 on MMSU perception, compared with 68.7 for Gemini-3.7 Flash, 63.8 for TAC, and 63.5 for Qwen3Omni-Captioner, while retaining strong reasoning performance (86.4). The attribute buildup in Figure 4 further localizes these gains, whereby adding individual schema attributes preferentially improves their associated perception tasks while leaving transcript-driven semantic tasks comparatively stable. This advantage is specific to fine-grained acoustic representation rather than general audio QnA alone. Frontier OmniLLMs remain weaker on fine-grained acoustics, timing, loudness, and phonetics, whether operating directly on audio or under the same *DtR* protocol (Table 2, right). Beyond MMSU, ITAD-S remains competitive on the broader MMAR and MMAU benchmarks, reaching 68.0 and 75.6, respectively. On the speaker-centric MSU-Bench, it achieves 82.2 overall (79.2 Tier 1; 86.6 Tier 2), outperforming the baselines and ITAD-E (Table 3). [1]

[1] For $\geq$ 30s clips, we chunk and use MOSS-0.9B (AI et al., 2026) solely for cross-chunk speaker consistency.

Table 2: **Inference and Reasoner Robustness. Left:** ITAD−S schemas under single-pass (SP) vs. multi-pass (MP) comprehensive decodes across text-only reasoners (top), and the MMSU tasks where MP helps most (bottom). **Right:** Perception fidelity. ITAD−S uses describe-then-reason; OmniLLM baselines process raw audio directly (left radar) or act as describers (right radar). ■ ITAD−S (ours), ■ Gemini 3.7, ■ Qwen3Omni

Text-only reasoner	SP	MP
ITAD−S → Gemini3	81.3	81.9
ITAD−S → Qwen3-Next	78.8	79.7
ITAD−S → Qwen3.8	80.5	80.9

Tasks (MMSU)	SP	MP
Pitch comparison	63.0	75.9
Vocal-range comp.	60.2	72.2
Emotional context	43.6	55.5
Accent ident.	59.6	69.3

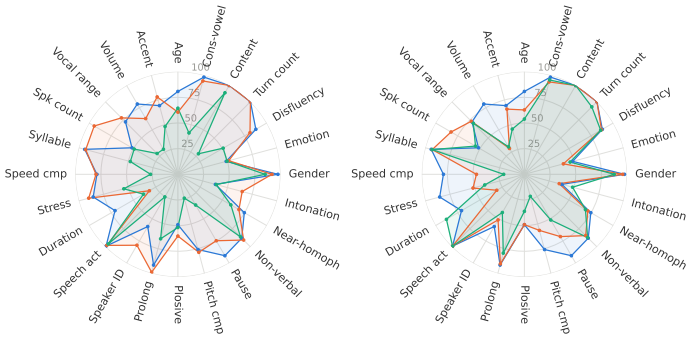

Table 3: **Speaker-centric understanding on MSU-Bench.** Evaluated on conversational multi-speaker scenarios in English/Mandarin. **Left:** Consistent with Table 1, ITAD−S uses a text reasoner. **Right:** Tier 1 (Identification) task accuracy. (AI: accent identification, AR: age recognition, GI: gender identification, ER: emotion recognition, SP: speaker profiling, SR: speaker retrieval, SVS: speaker-specific viewpoint summarization, SC: speaker counting, SV: speaker verification) OmniLLM baselines process raw audio directly (left radar) or act as describers (right radar). ITAD−S's trace (blue) is superior across both, demonstrating robust speaker-centric perception.

System	T1	T2	Avg
Kimi-Audio	41.0	47.0	43.0
Qwen2.5-Omni	19.0	21.0	19.0
StepAudio2	44.0	46.0	44.0
AudioFlamingo-3	40.0	38.0	39.0
MiMoAudio	52.0	64.0	56.0
Gemini-3.7 Flash	70.7	84.8	76.4
Qwen3-Omni	69.9	83.2	75.3
ITAD−E	<u>78.2</u>	**87.3**	<u>81.9</u>
ITAD−S	**79.2**	<u>86.6</u>	**82.2**



Inference and Reasoner Robustness. We ablate IIO decoding and the text reasoner in Table 2, left. Collapsing multi-pass (MP) inference, where each task conditions on the accumulated description, into a single (SP) decode incurs a modest MMSU drop (81.9 → 81.3). However, MP yields larger gains on perceptual tasks such as pitch (+12.9),vocal range (+12.0), andaccent identification (+9.7).Holding the ITAD−S schema fixed and swapping the reasoner from Gemini3.5-Flash with Qwen3-Next or Qwen3.8-Omni retains similar MP performance (79.7 and 80.9, respectively). In contrast to TAC (Kumar et al., 2026), where scaling the reasoner improved results, this relative robustness to reasoner choice suggests that ITAD−S's detailed schema already exposes much of the perceptual information needed for downstream reasoning. Together, these results show that IIO's dynamic control surface can improve fine-grained acoustic inference through intermediate conditioning, while the resulting schema remains effective across different text-only reasoners.

4.2 Generation: The Description as Paired Supervision for Audio Synthesis

Our second validation tests whether the structured perceptual representation is sufficiently grounded to be *renderable* back into audio. We use ITAD−S to relabel in-the-wild audio at scale, producing structured timestamped captions as paired supervision for a text-to-audio diffusion model. If the generated schemas faithfully capture the underlying scene, conditioning on its event timeline should reconstruct both *what* occurs and *when* it occurs (Figure 5).

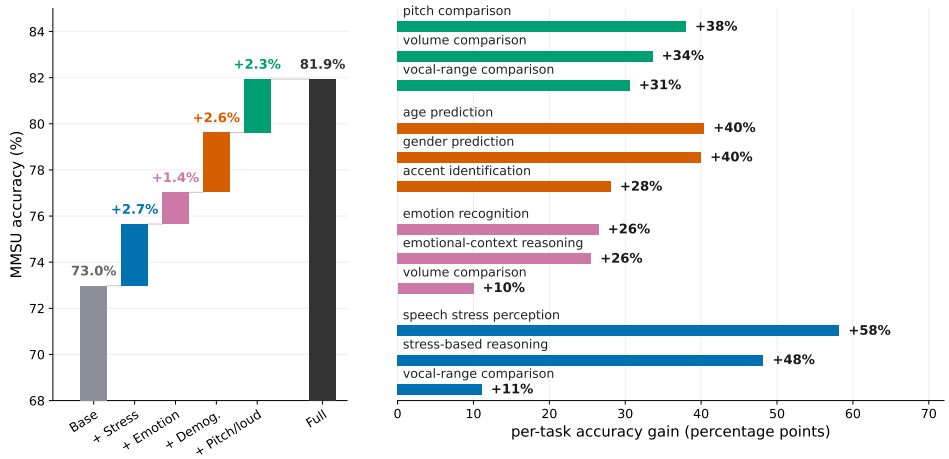

Figure 4: **Impact of individual IIO tasks. (left)** Building ITAD−S's description one attribute at a time, from a transcript-only base (who spoke, the words, and their timestamps). **(right)** For each attribute, the top-3 it lifts most, the gains land on the tasks that attribute governs and leave the transcript carried semantics tasks flat. Bar colour marks the added attribute (blue: stress and intonation, pink: emotion, orange: voice profile(demographics), green: pitch and loudness).

Table 4: **Structured captions generated by ITAD−S provide renderable supervision for audio synthesis.** Temporal-grounding cells report *with*|*without* input timestamps. PE and FLAM measure mean localizer-peak offset in seconds; PE-AV is a text-free frame-wise semantic distance between original and reconstructed audio (real-vs.-real = 0). Gemini-3-Pro judgments are on a 1–5 scale.

System	ITAD−S-as-judge							Temporal grounding			Gemini-3-Pro judge	
	Gen. ↑	Age ↑	Acc. ↑	Emo. ↑	Ev.R. ↑	Sp.R. ↑	WER ↓	PE ↓	FLAM ↓	PE-AV ↓	vs. real ↑	vs. cap. ↑
PicoAudio2 (Zheng et al., 2026)	0.44	0.35	0.24	0.28	0.27	0.09	1.15	2.47|2.77	3.16|3.50	0.52|0.51	1.39	1.61
MiDashengLM-Gen (Sun et al., 2026)	0.54	0.36	0.21	0.36	0.33	0.09	1.39	3.57|—	3.78|—	0.44|—	1.54	1.60
Seed Audio 1.0 (ByteDance Seed, 2026)	0.94	**0.89**	**0.71**	0.74	0.78	0.90	0.14	2.16|—	2.62|—	0.35|—	3.53	4.05
Ours (ITAD)	—	—	—	—	**0.80**	**0.94**	**0.12**	1.67|2.47	**2.13**|2.77	0.35|0.35	3.38	4.01
Ours (ITAD−S)	**0.99**	0.84	0.69	**0.77**	**0.80**	0.93	0.14	**1.54**|2.93	2.14|3.16	**0.30**|0.33	**3.95**	**4.52**

We condition the generator on the schema of held-out clips and localize events in the synthesized audio using two frame-level audio-text localizers, FLAM (Wu et al., 2025) and PE-A-Frame (Vyas et al., 2026). Temporal error is measured as the mean absolute offset between localizer peaks in the original and reconstructed audio (Table 4). Our generator provides the strongest temporal grounding, achieving 1.54s PE and 2.14s FLAM error, compared with 2.47s and 3.16s for the explicitly timestamp-conditioned PicoAudio2 (Zheng et al., 2026). Removing timestamps from the ITAD−S captions substantially degrades localization (1.54 → 2.93s PE; 2.14 → 3.16s FLAM), thereby demonstrating that the generator uses the predicted temporal structure.

Furthermore, we compare the recovered fields against the conditioning description using ITAD−S-as-judge. The reconstructed audio closely preserves speech content and speaker attributes, reaching 0.99 gender accuracy, 0.84 age accuracy, 0.69 accent accuracy, 0.77 emotion accuracy, 0.93 speech recall, and 0.14 WER. An audio-native Gemini-3-Pro judge rates our audio highest for both resemblance to the original (3.95) and adherence to the caption (4.52). We match the strongest commercial baseline, Seed Audio 1.0 (ByteDance Seed, 2026), on speaker attributes, whilst trailing on event timing and on both judgments. Additional details are in Appendix G.

In Figure 6, we demonstrate that ITAD−S's structured schema exposes explicit attributes as generation-time controls. For a fixed sentence, we vary only the emphasized word and its intensity in the ITAD−S' caption. Without training a dedicated stress-control model, the synthesized stress peak follows the commanded word and increases with the requested emphasis level. Thus, paralinguistic attributes elicited for perception also become explicit control channels when the same representation is used as supervision for generation.

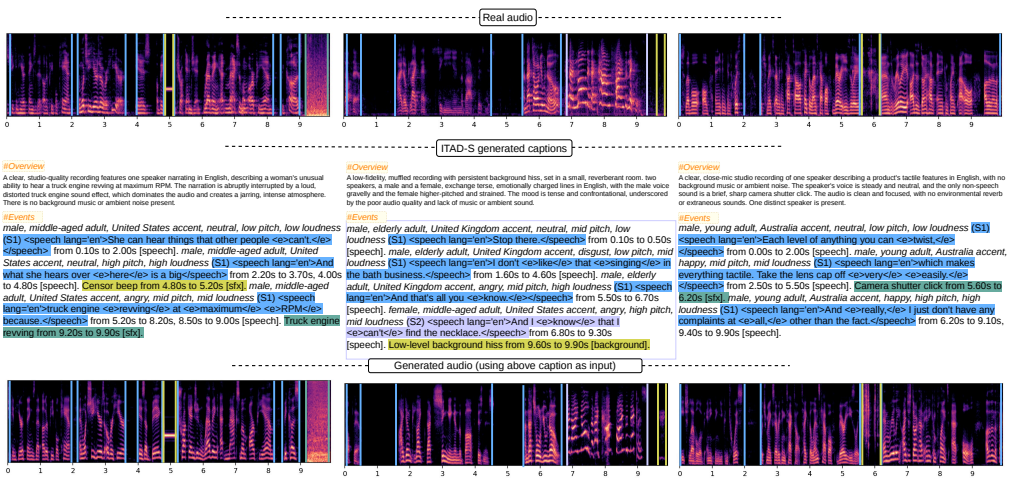

Figure 5: **ITAD−S-conditioned reconstruction.** We caption real audio (top) with ITAD−S (middle), then condition the trained text-to-audio generator on it to synthesize new audio (bottom). The reconstruction reproduces speech and non-speech events, including background elements like a censor beep, engine rev, camera click, at times aligned with the real audio.

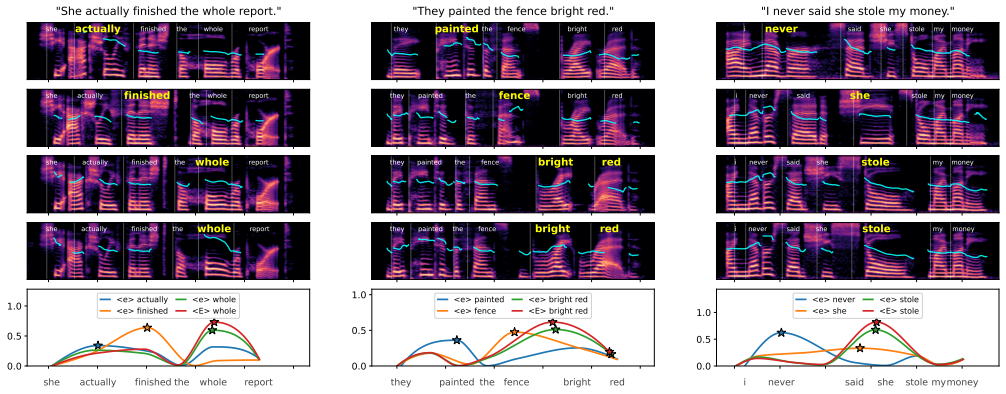

Figure 6: **Word-level emphasis as a renderable control channel.** Each column synthesizes the same sentence while varying the emphasized word in the ITAD−S caption (yellow). The bottom two rows increase emphasis from mild (<e>) to strong (<E>) while the cyan curve shows the F0 contour. Bottom panels show WhiStress (Yosha et al., 2025) stress probability over time, with ⋆ marking the commanded word. The detected stress peak follows the targeted word and scales with emphasis intensity, demonstrating zero-shot emphasis control without a dedicated stress-control model.

5 Conclusion

We introduce the Iterative Timestamped Audio Describer (ITAD) and its speech-centric extension, ITAD−S, for constructing queryable, temporally grounded acoustic representations from in-the-wild audio. Our framework bootstraps granular, speaker-attributed timelines at scale through multi-view weak supervision, models them through the flexible Instruction-Input-Output (IIO) completion format, and extends them with paralinguistic attributes through sample-efficient latent capability elicitation. The resulting unified model provides a structured interface between perception, reasoning, and generation. Paired with frozen text-only LLMs, it enables strong spoken audio understanding. Used as supervision for text-to-audio generation, it supports faithful reconstruction and explicit control over event timing, speaker attributes, and word-level emphasis. Overall, our results demonstrate that a high-fidelity, structured, temporally grounded acoustic timeline can function as both an *interpretable* reasoning-ready representation of audio and a *renderable* control interface for generation.

AI USE STATEMENT

The authors used generative AI to assist with language editing of the manuscript. We have reviewed and take responsibility for the final content of this work, including text produced with the aid of generative AI. All scientific content and conclusions were produced by the authors.

REPRODUCIBILITY STATEMENT

Our two-stage method is described in Section 3, with architecture and training details (backbone, LoRA rank and adapted modules, optimizer, step counts for `ITAD` and `ITAD-S`) in Appendix B. The time-registered IoU matching is formalized in Eq. 1–3 (Figure 2). The IIO format, and the per-task mixture and inference regimes in Section 3.1, Figure 3, and Appendix D; The specialist elicitation, probing setup, and per-attribute gold datasets are detailed in Section 3.2 and Appendix C. The text-to-audio validation (generator, evaluation set, metrics, baselines, and judge) and all evaluation protocols and judge prompts are provided in Appendices G and H. Our backbone (Qwen3-Omni-Captioner), the specialist gold corpora (IEMOCAP (Busso et al., 2008), RAVDESS (Livingstone & Russo, 2018), MELD (Poria et al., 2019), Common Voice (Ardila et al., 2020)), and all benchmarks are publicly available. Our in-the-wild corpus uses proprietary licensed audio sources, instead we provide complete time-registered IoU matching to reproduce our results with alternative audio sources.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 7319–7328, 2021.

MOSI AI, Donghua Yu, Zhengyuan Lin, Hanfu Chen, Chen Yang, Yiyang Zhang, Jingqi Chen, Ke Chen, Liwei Fan, Yi Jiang, et al. Moss transcribe diarize technical report. *arXiv preprint arXiv:2601.01554*, 2026.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pp. 4218–4222, 2020.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5185–5198, 2020.

Albert S Bregman. *Auditory scene analysis: The perceptual organization of sound*. The MIT press, 1990.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

ByteDance Seed. Seed audio 1.0: Audio creation model. `https://seed.bytedance.com/en/seedaudio1_0`, 2026.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. Aha-bench: Benchmarking audio hallucinations in large audio-language models. *Advances in Neural Information Processing Systems*, 38, 2026.

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

Elizabeth Donoway, Hailey Joren, Arushi Somani, Henry Sleight, Julian Michael, Michael Deweese, John Schulman, Ethan Perez, Fabien Roger, and Jan Leike. Quantifying elicitation of latent capabilities in language models. *Advances in Neural Information Processing Systems*, 38: 150412–150443, 2026.

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

Xuelong Geng, Kun Wei, Qijie Shao, Shuiyun Liu, Zhennan Lin, Zhixian Zhao, Guojian Li, Wenjie Tian, Peikun Chen, Yangze Li, et al. Osum: Advancing open speech understanding models with limited resources in academia. *arXiv preprint arXiv:2501.13306*, 2025.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 52423–52443, 2026.

Jeff Hwang, Moto Hira, Caroline Chen, Xiaohui Zhang, Zhaoheng Ni, Guangzhi Sun, Pingchuan Ma, Ruizhe Huang, Vineel Pratap, Yuekai Zhang, et al. Torchaudio 2.1: Advancing speech recognition, self-supervised learning, and audio processing components for pytorch. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–9. IEEE, 2023.

Can Jin, Hongwu Peng, Mingcan Xiang, Qixin Zhang, Xiangchi Yuan, Amit Hasan, Ohi Dibua, Yifan Gong, Yan Kang, and Dimitris N Metaxas. Dtop-p moe: Sparsity-controlled dynamic top-p moe for foundation model pre-training. *arXiv preprint arXiv:2512.13996*, 2025.

Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu. Speechcraft: A fine-grained expressive speech dataset with natural language description. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1255–1264, 2024.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019.

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

Chun-Yi Kuan and Hung-yi Lee. Can large audio-language models truly hear? tackling hallucinations with multi-task assessment and stepwise audio reasoning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Sonal Kumar, Prem Seetharaman, Ke Chen, Oriol Nieto, Jiaqi Su, Zhepei Wang, Rithesh Kumar, Dinesh Manocha, Nicholas J Bryan, Zeyu Jin, and Justin Salamon. Tac: Timestamped audio captioning. *arXiv preprint arXiv:2602.15766*, 2026.

Steven R. Livingstone and Frank A. Russo. The Ryerson audio-visual database of emotional speech and song (RAVDESS). *PLoS ONE*, 13(5):e0196391, 2018.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *Advances in Neural Information Processing Systems*, 38, 2026.

Max Morrison. torchcrepe, 2023. URL https://github.com/maxrmorrison/torchcrepe.

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

Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95, 2005.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 527–536, 2019.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Xingwei Sun, Heinrich Dinkel, Gang Li, Jiahao Mei, Yadong Niu, Zerui Han, Yuepeng Jiang, Jiahao Zhou, Lichun Fan, and Jian Luan. Midashenglm-gen: Unified audio scene generation via llm-driven autoregressive flow matching. *arXiv preprint arXiv:2608.11804*, 2026.

Kim Sung-Bin, Oh Hyun-Bin, Lee Jung-Mok, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *International Conference on Learning Representations*, volume 2025, pp. 24244–24271, 2025.

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

Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matthew Le, Piotr Dollár, Christoph Feichtenhofer, et al. Pushing the frontier of audio-visual perception with large-scale multimodal correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30172–30182, 2026.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. Blsp-emo: Towards empathetic large speech-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19186–19199, 2024.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2026, pp. 31374–31410, 2026a.

Dingdong Wang, Shujie Liu, Tianhua Zhang, Youjun Chen, Jinyu Li, and Helen Meng. Emotion-thinker: Prosody-aware reinforcement learning for explainable speech emotion reasoning. *ICLR*, 2026b.

Shuai Wang, Zhaokai Sun, Zhennan Lin, Chengyou Wang, Zhou Pan, and Lei Xie. Msu-bench: Towards understanding the conversational multi-talker scenarios. *Interspeech*, 2026c.

Megan Wei, Deepali Aneja, Jiaqi Su, Yunyun Wang, Haonan Chen, and Zeyu Jin. Do speech emphasis models generalize across languages and emotions? *arXiv preprint arXiv:2606.27717*, 2026.

Yusong Wu, Christos Tsirigotis, Ke Chen, Cheng-Zhi Anna Huang, Aaron Courville, Oriol Nieto, Prem Seetharaman, and Justin Salamon. Flam: Frame-wise language-audio modeling. *arXiv preprint arXiv:2505.05335*, 2025.

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

Iddo Yosha, Dorin Shteyman, and Yossi Adi. Whistress: Enriching transcriptions with sentence stress detection. *arXiv preprint arXiv:2505.19103*, 2025.

Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, et al. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025.

Zihao Zheng, Zeyu Xie, Xuenan Xu, Wen Wu, Chao Zhang, and Mengyue Wu. Picoaudio2: Temporal controllable text-to-audio generation with natural language description. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 14507–14511. IEEE, 2026.

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

Summary of Appendix

- Section A: Limitations and Future Work
- Section B: Implementation Details
- Section C: Elicitation: Per-Layer Probing
- Section D: IIO Tasks
- Section E: Segment- and pause-level localization
- Section F: Extending the Describer to Audio-Visual Reasoning
- Section G: Text-to-Audio Model and Temporal Grounding Setup
- Section H: Prompts

A Limitations and Future Work

Writing perception as an explicit textual description makes `ITAD-S`'s errors readily interpretable, tracing back to either an incorrect recognition of an existing attribute or a lack of schema coverage. Currently, `ITAD-S`'s schema prioritizes spoken attributes over non-speech and music events, a limitation reflected in benchmarks such as MMAR and MMAU. Interestingly, our current schema format enforces a temporal resolution tradeoff. Repeated events are merged into one span (for example, `<speech>...</speech> from 1.4s to 3.8s, 5.4s to 6.7s`) with comma-separated timestamps. While this fine temporal granularity helps `ITAD-S` act as a crucial reference for tasks like pause detection, disfluency detection, duration estimation etc. (see Table 2 (right), and Appendix E), it occasionally merges turns when consecutive voices are similar, causing the model to under-count unique speakers.

It is important to note that these bottlenecks reflect coverage limits rather than our architectural constraints. Our IIO proposal is inherently extensible, and new perceptual axes as well as upgraded teachers can be seamlessly integrated via teacher-forced bootstrapping. More broadly, human-designed schemas will eventually hit a coverage ceiling, oversight leading to bottleneck. In future, we aim to let descriptions evolve autonomously via reinforcement learning, using frozen text reasoners to reward the extraction of task-demanded cues.

B Implementation details

`ITAD` and the individual speech specialists are independently trained as rank-512, attention-only LoRA adapters over the same frozen Qwen3-Omni-30B captioner. At inference, the specialists can therefore be hot-loaded over this shared backbone to form the modular `ITAD-E` cascade. The unified `ITAD-S` is instead trained as an additional LoRA over the Stage 1 `ITAD` weights.

Stage 1. We train ITAD on approximately two million licensed, in-the-wild audio clips, each up to 30 seconds long. Each clip is paired with a fused target timeline constructed using the multi-view fusion procedure described in Section 3.1. We combine $K = 3$ specialized weak teachers: (1) a TAC-style sound event detector (Kumar et al., 2026) for non-speech event spans and types; 2) a TAC-inspired diarizer (Kumar et al., 2026) for speech turns and speaker identities; and (3) the Qwen3-Omni captioner (Xu et al., 2025) for transcripts, voice descriptions, and scene overviews. We additionally apply wav2vec2-based (Baevski et al., 2020) forced alignment to obtain word-level timestamps. The per-type overlap gate τ_k of Eq. 2 is set to 0.2 IoU for sound-effect events and 0 for speech and other non-speech events (rely on the order-preserving alignment with a midpoint-proximity fallback). A separate 0.5 IoU gate de-duplicates overlapping event descriptions. We selected these values by manual inspection of the merged outputs on a small validation set of clips. The resulting targets are decomposed into the IIO tasks described in Section 3.1, with training samples weighted proportionally to their speaker count. We optimize ITAD for 10,000 steps with a learning rate of 1.6×10^{-3}, a 1,000-step linear warmup, and cosine decay.

Stage 2. We train each speech specialist on fewer than 10,000 labeled examples, using task-specific teacher-forced context so that the adapter predicts only its designated paralinguistic attribute. For word stress and intonation, we use the multilingual A560E dataset (Wei et al., 2026), spanning 10 languages, and are conditioned on the transcript, incorporating per-word relative pitch extracted via *torchcrepe* (Morrison, 2023). For emotion, we use IEMOCAP (Busso et al., 2008), RAVDESS (Livingstone & Russo, 2018), and MELD (Poria et al., 2019), predicting utterance-level emotion directly from the audio. For voice profile (gender, age, and accent), we use silver labels obtained from a per-speaker Qwen3-Omni probe over the licensed corpus, supplemented with a balanced Common Voice (Ardila et al., 2020) accent shard to mitigate the US-accent skew of the silver labels, conditioned on the caption. Finally, relative pitch and loudness are derived from DSP measurements (CREPE F0 and RMS energy) ranked within each clip and conditioned on the temporal event list.

Once trained, these specialists are concurrently hot-loaded onto the ITAD backbone to form ITAD−E, which is subsequently used to re-annotate the complete Stage 1 corpus. A two-pass LLM rewriting step, using Qwen3-Next-80B, then rectifies the overview with the per-event descriptors from the hot-loaded specialists, for example to keep speaker count, gender, and accent consistent.

We initialize ITAD−S from the Stage 1 ITAD weights and train an additional rank-512, attention-only LoRA on the enriched corpus. The elicited speech attributes are introduced as additional IIO tasks and mixed with the original Stage 1 objectives, preserving the same IIO fan-out used by ITAD.

We train ITAD−S for 2,000 steps with a 100-step linear warmup and cosine decay. The resulting model consolidates the modular predictions of ITAD−E into a single adapter while retaining the structural capabilities learned by ITAD.

C Elicitation: Per-Layer Probing

Section 3.2 hypothesizes that the Qwen3-Omni backbone already encodes paralinguistic information internally, while our attention-only LoRA specialists primarily improve its accessibility to late-stage generation rather than learning new perceptual features. We examine this hypothesis through layer-wise linear probing before and after specialist adaptation.

Per-layer probe. For each paralinguistic attribute, we fit an L_2-regularized logistic probe independently at each of the model's 49 hidden layers. Representations are extracted at the final token position corresponding to the attribute's generation step. We report 5-fold cross-validated AUC, with macro one-vs-rest averaging for multiclass attributes such as accent.

Three consistent patterns emerge (Table 5). First, every evaluated trait exhibits high linear decodability within the frozen backbone (peak AUCs between 0.89 and 0.99). Second, across the early and middle transformer blocks (depths 0–24), the representation quality of the backbone and the specialist are virtually identical (e.g., word stress AUC is 0.88 for the base versus 0.89 for the specialist), confirming that the adapter preserves early acoustic processing. Third, the adapter's structural impact is strictly localized to the late layers (36–48), where it maximizes performance within the available headroom. This relative gain is most pronounced for relational traits that the base model struggles to articulate (e.g., relative loudness and pitch, yielding a $+0.10$ to $+0.12$ AUC lift).

Table 5: The backbone (Qwen3Omni-captioner) possesses latent perception; the specialist isolates late-stage readouts. *Early*/*Late*: mean AUC over depths 0–24 / 36–48, backbone vs. specialist. *Lift*: the specialist's late-stage gain.

Trait (granularity)	Backbone peak	Early (base / spec)	Late (base → spec)	Lift
Relative loudness (segment)	0.89	0.79 / 0.81	0.87 → 0.99	+0.12
Relative pitch (segment)	0.90	0.80 / 0.82	0.89 → 0.99	+0.10
Word stress (word)	0.93	0.88 / 0.89	0.91 → 0.99	+0.08
Accent (speaker)	0.99	0.97 / 0.97	0.99 → 1.00	+0.01

Table 6: The specialist mitigates output mode collapse without altering early perception. Accuracy of the emitted textual attribute, frozen backbone vs. its specialist, on clips with ground-truth labels (word stress is scored as the top stressed word against Whistress (Yosha et al., 2025)).

Attribute (metric)	Frozen backbone	+ specialist
Word stress (top-word accuracy)	0.60	0.81
Accent (exact, 13 classes)	0.37	0.62
Accent (region)	0.61	0.91
Emotion (5 classes)	0.15	0.59

For traits already near the performance ceiling internally (e.g., accent at $+0.01$ lift), the specialist functions primarily to calibrate the final output distribution rather than enhance latent separability.

Mitigating output collapse and bias. High internal decodability does not inherently translate to accurate generation. Post-training alignment heavily biases the decoding trajectory, causing the base model's textual outputs to exhibit mode collapse toward majority classes. Table 6 contrasts the generated attribute accuracy of the frozen backbone against the specialist on a ground-truth evaluation set.

The baseline generation errors are highly systematic. Accent predictions overwhelmingly default to "United States" (emitted for 42% of samples compared to a 9% true prior), and emotion predictions degenerate to "neutral" (76% of samples, despite "neutral" not appearing as a valid ground-truth class). Concurrently, word stress exhibits a severe false-positive rate (1.5 stressed words per clip at 0.21 precision).

The specialist adapter effectively mitigates this generation bias, substantially recovering the true attribute distributions without altering the underlying perceptual manifold (consistent with the identical early-layer AUCs). Furthermore, the specialist's residual accent errors primarily occur between geographically adjacent regions (e.g., Canada → United States, Ireland → United Kingdom, New Zealand → Australia, Kenya → Nigeria). Consequently, strict 13-class exact match accuracy underestimates the model's true perceptual fidelity; evaluating at the broader regional level increases the specialist's accuracy to 0.91, compared to 0.61 for the frozen backbone.

D IIO TASKS

The unified student model is trained using the extension of our proposed Instruction-Input-Output (IIO) format. During training, each example is formatted as a single completion task, sampled according to fixed probabilities. The mixture comprises per-attribute tasks which describe isolated paralinguistic features and one `comprehensive` task designed to output the entire structured description at once. The comparative attribute specialists always receive the speaker turns and transcript as teacher-forced input. Conversely, the `caption` and `comprehensive` tasks take no text input and decode directly from the raw audio (Table 7). To establish single-pass capabilities, we append the `comprehensive` task (empty `INPUT`) at a weight of 1.0 and reduced the `caption` task weight to 0.5. The comprehensive task adds single-pass capability while retaining the cascaded multi-pass inference regime.

Table 7: **IIO tasks during training.** Each training sample is formulated as a single completion task, with weights denoting their respective sampling probabilities. The comparative speech specialists are conditioned on the speaker turns and transcript as teacher-forced inputs.

Task	Input	Weight	Output
comprehensive	none (audio)	1.0	full description (all sections)
caption	none (audio)	0.5	overview + events with transcript
stress	transcript turns	0.5	per-word stress and intonation
emotion	transcript turns	0.5	per-turn emotion
voice profile	transcript turns	0.3	per-speaker gender, age, accent
acoustics	transcript turns	0.3	per-segment pitch and loudness

Table 8: **Segment- and pause-level localization on conversational speech.** Median onset error (ms, ↓) and boundary F1 at 100/200 ms (↑). *Count*: fraction of windows where `ITAD` proposes the correct segment count.

System	Onset (ms)	F1@100	F1@200	Count
Qwen3-FA aligner Shi et al. (2026)	**23**	**0.784**	0.823	n/a
wav2vec2 aligner Hwang et al. (2023)	72	0.608	0.752	n/a
`ITAD` (ours)	49	0.774	**0.929**	84%

At inference, the dynamic IIO surface allows the model to operate in two distinct regimes. The multi-pass approach executes a cascaded sequence of five decodes. First, the `caption` task processes the raw audio to generate the scene overview and the timestamped events list, including the verbatim transcript. The predicted speaker turns from this initial pass are subsequently fed back as the `INPUT` conditioning for the four specialist decodes (`stress`, `emotion`, `voice profile`, and `acoustics`). This iterative teacher-forcing ensures all fine-grained paralinguistic descriptors remain structurally aligned to the same master transcript. Conversely, the single-pass approach utilizes only the `comprehensive` task. By reading the raw audio with an empty `INPUT` IIO, it emits the complete structured description, spanning all non-speech events as well as speech attributes, in a single, continuous output block. Both inference approaches yield the identical structured schema.

E SEGMENT- AND PAUSE-LEVEL LOCALIZATION

`ITAD`'s training schema records a turn broken by a pause as a comma-separated list of active spans (for example, `<speech>...</speech> from 1.4s to 3.8s, 5.4s to 6.7s`). We test whether this supervision produces accurate segment- and pause-level localization on real conversational speech, and how it compares to dedicated aligners.

Specifically, we evaluate through the `caption` task (Appendix D) wherein we mark segment boundaries in the reference transcript with the delimiter (\\), and recover each segment's predicted timestamps. A segment is a run of speech bounded by a pause, whether a change of speaker or the same speaker pausing mid-utterance.

We use the Buckeye corpus of conversational American English Pitt et al. (2005) (eight development speakers, 50 recordings). We define a segment as a run of target-speaker words bounded by a silence of at least 0.1 s or an interviewer turn (provided in the dataset), and pack consecutive segments into windows of at most 20 s yielding 1,349 windows and 4,947 segments.

For reference we run two specialized alignment models on the same windows, the Qwen3 aligner Shi et al. (2026) and a wav2vec2-based aligner Hwang et al. (2023). `ITAD` localizes segment onsets to a 49 ms median boundary error and attains the highest boundary agreement at a 200 ms tolerance (0.93), the most consistent of the three systems and competitive with models built solely for alignment (Table 8, Figure 7). Importantly, `ITAD` itself proposes the segmentation, recovering the correct segment count 84% of the time, whereas the aligners require this as input.

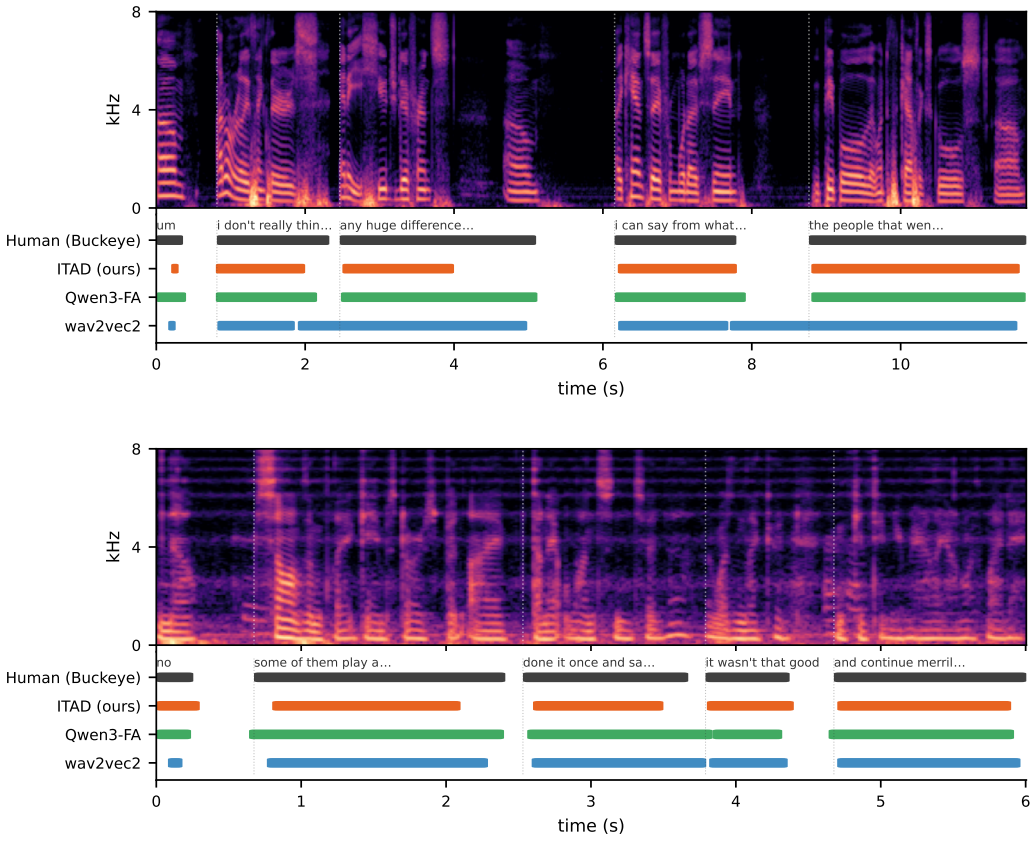

Figure 7: **Segment localization on conversational speech.** The segments from the human annotation, ITAD, the Qwen3-FA aligner, and the wav2vec2-based aligner share the time axis. ITAD's segment onsets track the human pause boundaries closely, while being competitive with the dedicated aligners. (Dotted lines mark the human segment onsets.)

F Extending the describer to audio-visual reasoning

The decoupled describe-then-reason paradigm is not limited to audio understanding but can readily be extended to serve as the acoustic foundation for an audio-visual describer. Following the TAC-V framework introduced by Kumar et al. (2026), we generate an integrated audio-visual caption by pairing our audio describer with an off-the-shelf visual grounder (Qwen3-VL-32B). Specifically, the audio describer outputs timestamped acoustic events and speech, while a visual shot list delineates scene boundaries. The visual grounder then links acoustic events to their on-screen sources, leveraging visual context to resolve acoustic ambiguities. Finally, a frozen text reasoner answers downstream queries utilizing this unified multimodal caption alone.

We evaluate on four audio-visual reasoning benchmarks: Daily-Omni (Zhou et al., 2025), WorldSense (Hong et al., 2026), Video-Holmes (Cheng et al., 2025) and AVHBench (Sung-Bin et al., 2025). Table 9 replicates the comparative setup of Kumar et al. (2026), contrasting native multimodal large language models (which process audio and video directly) against the cascaded describe-then-reason approach. The *VLM + Qwen3* ablation establishes a baseline using visual-only captions devoid of acoustic descriptions. Substituting ITAD-S improves the cascade on all three reasoning benchmarks ($+1.9$, $+0.4$, $+5.9$ over TAC-V) and beats every native model on Daily-Omni and Video-Holmes. On AVHBench, ITAD-S-V leads or matches TAC-V on the three *discriminative* subtasks (audio-driven video hallucination AVH, video-driven audio hallucination VAH, and audio-visual matching AVM) but trails on the *generative* captioning subtask (AVC). AVC evaluates a generated caption against a *single* human reference using METEOR. Interestingly, we find that even a faithful paraphrase of the reference itself attains only $\sim$35 METEOR. It is evident that word-

Table 9: **Audio-visual reasoning performance.** Our ITAD-S-V pipeline substitutes ITAD-S for the prior TAC audio describer, keeping describe-then-reason pipeline constant. Daily-Omni, World-Sense, and Video-Holmes report multiple-choice accuracy. AVHBench reports accuracy for AVH/VAH/AVM and METEOR for the AVC subtask. Best per row in bold.

	Native MLLM		Describe-then-reason				
Benchmark	Model	Score	VLM + Qwen3	TAC-V + Qwen3	TAC-V + Gemini3	ITAD-S-V + Qwen3	ITAD-S-V + Gemini3
Daily-Omni	Qwen3-Omni	76.2	51.5	72.9	77.9	77.2	**79.8**
	Gemini 2.5 Flash	72.7					
	OmniVinci	66.5					
World-Sense	Gemini 2.5 Pro	**65.1**	37.4	45.7	58.6	55.8	59.0
	OmniVinci	48.2					
Video-Holmes	Qwen3-Omni	57.3	45.6	47.7	59.2	58.4	**65.1**
AVHBench (AVH)	PandaGPT	58.5	70.8	79.8	81.7	82.0	**82.2**
AVHBench (VAH)	PandaGPT	61.3	51.8	76.1	76.6	**76.9**	76.4
AVHBench (AVM)	OneLLM	60.1	50.5	56.7	61.6	72.6	**77.5**
AVHBench (AVC)	Video-LLaMa	14.0	12.9	**22.6**	20.6	14.1	13.8

ing differences alone therefore forfeit roughly two-thirds of the achievable score, compressing the absolute range for every system. We also noticed that the per-clip METEOR is high exactly when our caption's opening subject and word choice happens to coincide with the reference and low otherwise. We believe the metric tracks *phrasing agreement*, not caption correctness. We form ITAD-S-V by substituting ITAD-S for TAC's audio describer while keeping the same visual grounder and the reasoner, so the improvement is only attributable to our generated audio description.

G TEXT TO AUDIO MODEL AND TEMPORAL GROUNDING SETUP

Generator. Derived from (Jin et al., 2025), our text-to-audio (T2A) architecture is a latent diffusion model that synthesizes 48 kHz audio conditioned on text captions. The base model is trained exclusively on structural ITAD captions, which consist of a scene overview and a timestamped event list devoid of speaker characteristics. A subsequent fine-tuning phase incorporates enriched ITAD-S captions (include speaker demographics and word-level emphasis tags), that enable explicit control over speaker attributes while maintaining the temporal fidelity (see Section 4.2). To evaluate the base T2A model, we deterministically ablate speaker metadata and emphasis tags from the enriched captions, retaining only the scene overview and timestamped events.

ITAD-S-as-judge. Beyond timing, we test whether the full schema survives synthesis with a describe–generate–re-describe approach. We re-caption the generated audio with ITAD-S and match the re-caption to the input caption field by field. We report accuracy for speaker gender, age, and accent and for emotion, recall for all events and for speech events, and transcript word error rate. A missing or mismatched attribute counts as an error, so every field is scored over the full clip set. This tests whether the commanded speaker attributes, events, and spoken content are recoverable from the generated audio, complementing the temporal and perceptual metrics below.

Temporal Grounding Evaluation. We construct a diverse evaluation suite of 300 held-out English clips encompassing six accents (United States, United Kingdom, Nigeria, India, Australia, and the Philippines). Each clip contains between three and ten distinct acoustic events, yielding 1,471 total events. To measure temporal fidelity without treating the generated text caption as a reliable ground truth, we utilize the original recording as a zero-distance reference. We employ FLAM (Wu et al., 2025) and PE-A-Frame (Vyas et al., 2026), two frame-level audio-text models, to compute semantic similarity scores for every short-time frame against a given text query, extracting the peak activation frame as the predicted temporal location. For each event, we report the absolute temporal offset (in seconds) between the localizer's peak in the generated audio and its corresponding peak in the original recording. This error is averaged exclusively over the events successfully localized in the original recording to ensure a reliable reference.

The peak-offset metrics localize each event against a text query. We additionally report a fully text-free measure: using PE-A-Frame's audio tower alone, we embed the original and generated audio into one 1024-d vector per 40 ms frame and report their mean per-frame cosine distance averaged over clips. The distance to original audio is zero, as reference. We believe this is a recall-heavy, global reconstruction measure. It ranks the systems identically to the peak-offset metrics and tracks the Gemini-3-Pro judge, supporting the view that the ITAD-S caption is a compact latent expressed in text from which the scene can be re-rendered.

Baselines. We benchmark against MiDashengLM-Gen (Sun et al., 2026), a recent text-to-audio scene generator and PicoAudio2 (Zheng et al., 2026), a model explicitly designed for timestamp control, evaluating it both with and without event timestamps to provide a clean intra-model ablation. Note that PicoAudio2 does not model speech. Finally, we evaluate Seed Audio 1.0 (ByteDance Seed, 2026), a highly capable commercial text-to-audio model prompted with the identical caption formatted as a scene overview followed by timestamped event lines.

H PROMPTS

For MMSU, MMAR, and MMAU, for the audio-visual benchmarks, and for the open-weight reasoners in our robustness study (Section 4.1), we use the describe-then-reason prompts of Kumar et al. (2026). Below we provide the two prompts specific to this work: the describe-then-reason wrapper for MSU-Bench, and the audio-native judge used to evaluate text-to-audio synthesis.

H.1 MSU-BENCH REASONER

We score MSU-Bench with its official evaluation harness (Wang et al., 2026c), using its multiple-choice prompt (the question, the answer options, and a fixed output-format instruction) verbatim. We replace the audio with ITAD-S's text description of the segment(s) referenced by each question, prepending it as shown in Figure 8. The reasoner has access to only the caption generated by our models.

Describe-then-reason prompt for MSU-Bench

```
Audio description:
{description}

{question}

Options:
{options}

[Output format]
On the very last line of your reply, output only:  Final answer:
A/B/C/D.
```

Figure 8: The MSU-Bench describe-then-reason prompt. {description} is ITAD-S's text description of the referenced segment(s). The output-format instruction is the official MSU-Bench harness prompt (Wang et al., 2026c) used verbatim.

H.2 TEXT-TO-AUDIO PERCEPTUAL JUDGE (GEMINI-3-PRO)

To conduct holistic perceptual evaluations, we employ Gemini-3-Pro as an audio-native judge. The model is queried greedily with one call per (system, clip) pair across the evaluation set. The judge evaluates systems across two rubrics. *vs. real* passes both the original recording and the generated candidate, prompting the model to rate their acoustic and thematic resemblance on a 1-5 scale. *vs. caption* passes the enriched ITAD-S text caption alongside the generated audio, prompting the model to evaluate how faithfully the candidate synthesizes the requested description. Since every system is judged against the *same* ITAD-S caption, this metric provides a direct, level-field cross-model comparison.

Text-to-audio judge: *vs. real* recording

```
You are an expert audio evaluator.  You are given two short audio
clips:  REFERENCE, an original real-world recording, and CANDIDATE,
audio from a text-to-audio model asked to reproduce the same
scene.  Rate how closely the CANDIDATE resembles the REFERENCE
on an integer scale 1-5:  5 = near-identical (same distinct sound
events at essentially the same times, matching speaker/voice
characteristics, scene, and ambience); 4 = strong (most events
present and roughly aligned; minor differences); 3 = moderate
(general scene recognizable but several events missing, mistimed,
or altered); 2 = weak (loose thematic connection only); 1 = no
resemblance.  Weigh, in order:  (a) the same events are present,
(b) their timing and order, (c) speaker/voice characteristics for
any speech (gender, age, accent, delivery), (d) overall scene,
ambience, and recording quality.  Reason briefly, then end with a
JSON object on its own final line:  {"rating":  <1-5>}.
```

Figure 9: Gemini-3-Pro audio-native judge prompt for the *vs. real* rubric.

Text-to-audio judge: *vs. caption*

```
You are an expert audio evaluator.  You are given a DESCRIPTION,
a structured text description of an intended audio clip (it
may contain # Overview / # Events sections, per-event speaker
attributes, timestamps of the form 'from Xs to Ys', a transcript
inside <speech>...</speech>, and inline emphasis markup <e>word</e>
/ <E>word</E>), and a CANDIDATE, audio produced by a text-to-audio
model conditioned on this description.  Rate how faithfully the
CANDIDATE realizes the DESCRIPTION on an integer scale 1-5:  5 =
faithful (all described events at the described times and order;
for speech, the words, described speaker attributes, and any
marked emphasis all match; scene and ambience match); 4 = strong
(most described content present and roughly aligned); 3 = moderate
(scene recognizable but several described events, attributes, or
timings missing or wrong); 2 = weak (loose connection only); 1 =
unrelated.  Weigh, in order:  (a) the described events are present,
(b) their timing and order, (c) for speech, the spoken content
and described speaker attributes and marked emphasis, (d) overall
scene, ambience, and quality.  Judge only against the DESCRIPTION;
do not penalize plausible details it leaves unspecified.  Reason
briefly, then end with a JSON object on its own final line:
{"rating":  <1-5>}.
```

Figure 10: Gemini-3-Pro audio-native judge prompt for the *vs. caption* rubric. Since every system is judged against the same `ITAD-S` caption, this yields a level-field cross-model comparison.