TAC: Timestamped Audio Captioning

Sonal Kumar[1,2,*] **Prem Seetharaman**[2,*] **Ke Chen**[2] **Oriol Nieto**[2] **Jiaqi Su**[2]

Zhepei Wang[2] **Rithesh Kumar**[3,†] **Dinesh Manocha**[1] **Nicholas J. Bryan**[2] **Zeyu Jin**[2]

Justin Salamon[2]

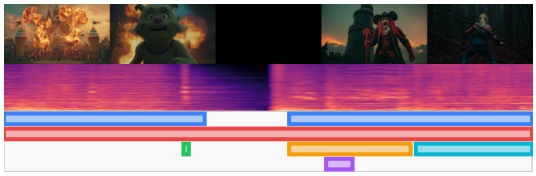
[1]University of Maryland, College Park, USA [2]Adobe Research, USA
[3]OpenAI, USA [*]Equal contribution [†]Work done while at Adobe
sonalkum@umd.edu, seethara@adobe.com

Abstract

Large Audio Language Models struggle to disentangle overlapping events in complex acoustic scenes, yielding temporally inconsistent captions and frequent hallucinations. We introduce Timestamped Audio Captioner (TAC), a model that produces temporally grounded audio descriptions at varying degrees of detail and resolution. TAC is trained with a synthetic data pipeline that constructs challenging and dynamic mixtures from real-world audio sources, enabling robust learning under realistic polyphonic conditions. Across event detection, dense captioning, and speaker diarization, TAC outperforms most competing methods - including dedicated diarization systems with a low hallucination rate and accurate temporal grounding. We also introduce TAC-V, an audio-visual pipeline to generate semantically rich audio-visual descriptions. We then show that TAC and TAC-V serves as a "semantic bridge" for a text-only reasoner: a simple TAC→LLM and TAC-V→LLM cascade achieves state-of-the-art scores on benchmarks for both audio (MMAU-Pro, MMSU, MMAR) and audio-visual (DailyOmni, VideoHolmes) understanding and reasoning respectively. We encourage readers to see detailed qualitative results on our demo page: `https://sonalkum.github.io/tacmodel/`.

1 Introduction

The pursuit of *audio general intelligence* is rapidly advancing with Large Audio-Language Models (LALMs), which promise to turn raw audio into rich semantic understanding for captioning, instruction following, and open-ended reasoning. Recent foundation models including SALMONN [Tang et al., 2024], Qwen2-Audio [Chu et al., 2024], GAMA [Ghosh et al., 2024], the Audio Flamingo series [Kong et al., 2024, Ghosh et al., 2025a, Goel et al., 2025], Audio-Thinker Wu et al. [2025a], Kimi-Audio Ding et al. [2025], and MiMo-Audio Zhang et al. [2025] have demonstrated impressive progress across speech, sound, and music understanding. Yet, when deployed on complex real-world auditory scenes with *overlapping* and *time-varying* events, these systems



Audio → TAC → **[music] Heroic brass fanfares and thunderous percussion from 0.0s to 3.8s, 5.4s to 10.0s. [sfx] Fire crackling and burning from 0.0s to 10.0s. Sudden burst of sound from 3.4s to 3.5s. [sfx] A group of people shouting in unison, expressing excitement from 5.4s to 7.7s. [sfx] Heavy object crashes down from 6.1s to 6.6s. [sfx] Rattling and clattering from a moving chain from 7.8s to 10.0s.**

Figure 1: Given only audio, TAC generates structured, timestamped descriptions of overlapping sound events. We visualize the timestamps produced by TAC as temporal lanes above. Colors indicate correspondence between text and temporal lanes.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

remain brittle. Even strong frontier models (e.g., Gemini 3 Pro [Team and Google, 2025]) often produce *global* captions that miss fine-grained temporal structure, confuse event boundaries, or hallucinate non-existent sounds – failure modes that recent benchmarks and analyses identify as central obstacles to reliable audio understanding [Kuan and Lee, 2025, Cheng et al., 2025b].

We argue that these failures reflect a fundamental *supervision mismatch* between continuous, high-density audio streams and the sparse language annotations used to train LALMs. Popular captioning datasets (e.g., AudioCaps [Kim et al., 2019], Clotho [Drossos et al., 2020]) typically provide a single caption for a 10–30 second clip. This results in *semantic collapse*: temporally distinct events are compressed into a short, clip-level summary, making it difficult for models to preserve causality and disentangle overlaps. Language priors can then dominate and yield hallucinations [Kuan and Lee, 2025, Cheng et al., 2025b]. Recent alignment efforts further suggest that grounding failures are systemic, and can be reduced only when training includes hard counterfactual negatives targeting fine-grained temporal reasoning [Cheng et al., 2025b]. These findings indicate that robust audio understanding requires bridging dense audio with *structured, temporally grounded* linguistic supervision.

We propose Timestamped Audio Captioner (TAC), a model trained to produce timestamped audio description (see Fig. 1). TAC produces captions paired with exact start and end times for every source in complex auditory scenes. Unlike prior LALMs which tackle broader understanding and reasoning Ghosh et al. [2025a], Goel et al. [2025], Ghosh et al. [2024], Team and Google [2025], Xu et al. [2025], TAC focuses on "what happens when" (e.g. sound event detection). We then cascade TAC with a "reasoner" (a text-only LLM), resulting in a "describe-then-reason" approach to multimodal understanding. From audio, TAC produces high-quality dense text captions that serve as evidence that the reasoner uses to answer questions. Finally, we extend this to audiovisual inputs by pairing TAC with an off-the-shelf VLM. Remarkably, we find that this simple cascade obtains state-of-the-art results on several multimodal understanding benchmarks. By decoupling the describer from the reasoner, we can scale the two components independently. We show that stronger reasoners give higher performance, even when given access to the same TAC descriptions.

While our architecture follows the standard LALM paradigm, our key insight is that the bottleneck lies in *supervision granularity*, not model complexity. We show that a principled data pipeline with dynamic supervision including a weighted loss (Eq. 1) and multitask prompt conditioning enables a standard backbone to surpass far larger models trained on orders of magnitude more data. Our contributions are: (i) **TAC:** an audio understanding model trained on a synthetic, multi-granular curriculum generated by a dynamic data generation pipeline, achieving SOTA results in audio captioning, sound event detection (SED) and speaker diarization tasks; (ii) **TAC-V:** an audio-visual extension obtained by pairing TAC with a vision–language model to produce dense audio-visual captions; and (iii) **Describe then reason:** dense captions from TAC(-V) are a semantic bridge for reasoning with text-only LLMs, yielding state-of-the-art performance on audio reasoning benchmarks (*MMAR* Ma et al. [2025a], *MMSU* Wang et al. [2025], *MMAU-Pro* Kumar et al. [2025]) and competitive results on *MMAU* Sakshi et al. [2025], as well as state-of-the-art or competitive audiovisual reasoning performance when combining TAC-V with a text-only LLM (*DailyOmni* Zhou et al. [2025], *VideoHolmes* Cheng et al. [2025a], *WorldSense* Hong et al. [2026], *AVHBench* Sung-Bin et al. [2025]).

2 Related Work

LALMs. Recent work in audio perception and understanding has shifted from task-specific models [Gong et al., 2021, Chen et al., 2023] to general-purpose generative systems. Works like LTU [Gong et al., 2024] and SALMONN [Tang et al., 2024] demonstrated that aligning audio encoders (e.g., Whisper, AudioMAE) with LLMs enables zero-shot speech and audio reasoning. Instruction-tuned models, such as GAMA Ghosh et al. [2024], Qwen-Audio [Chu et al., 2023] and Audio Flamingo series[Kong et al., 2024, Ghosh et al., 2025a, Goel et al., 2025], have scaled approach, achieving impressive performance by embedding audio directly into the context of an LLM. AudioChat Chen et al. [2026] enables audio foundation models to generate, edit, and understand complex "audio stories" (multi-speaker, multi-source scenes) by simulating realistic training data with LLM agents and training with Audio Transfusion Forcing. However, these models often falter in "cocktail party" scenarios involving overlapping sound events. Even strong frontier models like Gemini 3 Pro [Team and Google, 2025] still remain prone to hallucinating events not present in the audio [Kuan and Lee,

2025]. We attribute this to the "global pooling" nature of their supervision, where temporal details are compressed into a single semantic vector. In contrast, TAC enforces a dense, time-aware alignment, enabling detailed reasoning.

Audio and Audiovisual Captioning. Automated Audio Captioning (AAC) has traditionally relied on human-annotated datasets like AudioCaps [Kim et al., 2019] and Clotho [Drossos et al., 2020], limited by fewer samples and "sparse" annotation—a single sentence for a 10–30 s clip. This forces models to learn correlations rather than causality. While dense captioning is well-explored in vision [Johnson et al., 2016], it remains nascent in audio due to the cost of dense timestamp annotation. Weakly-supervised approaches like WavCaps [Mei et al., 2024] scale up using noisy metadata but lack precise temporal boundaries for SED. Although AudioSet-Strong Hershey et al. [2021] offers timestamped event labels and TACOS Primus et al. [2025] targets temporal alignment with human-annotated clips, their focus is atomic classification and free-text SED, not dense descriptions. In the audiovisual domain, Video-SALMONN 2 Tang et al. [2025] extends SALMONN with caption-enhanced training, while Omni-Captioner Ma et al. [2025b] and AuroraCap Chai et al. [2024] target omni-modal and video-level dense perception. TAC addresses annotation scarcity by synthesizing a curriculum of dense, temporally-precise captions bridging simple tagging and complex storytelling.

Synthetic and LLM-based Data Generation for Audio. Recent works use LLMs to generate QA pairs or captions from audio metadata. GAMA [Ghosh et al., 2024] and Audio Flamingo 2/3 [Ghosh et al., 2025a, Goel et al., 2025] utilize GPT-4 for complex QA and reasoning chains, while ReCLAP [Ghosh et al., 2025b] augments training data by rewriting captions to emphasize acoustic characteristics. These focus on global clip-level understanding but lack fine-grained temporal grounding. Scaper [Salamon et al., 2017] mixes isolated events for closed-vocabulary SED; FLAM [Wu et al., 2025b] mixes audio to train a contrastive frame-wise discriminator; and Hive [Li et al., 2026] synthesizes semantically consistent mixtures for query-based universal sound separation. Unlike these objectives, TAC generates multi-granular, open-vocabulary caption supervision with exact timestamps and jointly sampled style, resolution, merge, and activity controls. In the audiovisual domain, AVoCaDo Chen et al. [2025a] generates temporally orchestrated captions using Gemini 2.5-Pro with GPT-4.1 filtering, while Omni-Detective Ma et al. [2025b] orchestrates multiple frontier models with up to 10 iterative rounds per sample. TAC-V instead uses a single TAC pass followed by a single VLM pass.

3 Methodology

We introduce TAC, a model designed to bridge the gap between low-level acoustic signals and high-level reasoning. This pipeline allows us to finetune a standard LALM [Chu et al., 2023] to achieve state-of-the-art dense captioning within just 5k training iterations over synthetic mixtures. The proposed methodology is depicted in Figure 2, and below we detail all its respective steps.

3.1 Dynamic Acoustic Mixer

While recent works scale model size to improve performance, we argue that the bottleneck lies in the *granularity* of supervision. Standard datasets provide a single "global" caption for a complex scene, forcing models to average out temporal details. To overcome this, we use a **Dynamic Acoustic Mixer** that generates infinite, highly-complex audio mixtures with synchronized ground truth at multiple levels of semantic resolution from single-source audio datasets.To increase the realism of the mixer, we define acoustic scenes via **Scene Templates** that specify the structural logic of an audio clip. A template T consists of a set of temporal constraints C and role bindings $R = \{r_{speech}, r_{music}, r_{sfx}, r_{bg}\}$. For example, a "Speech over Music in Indoor Environment"

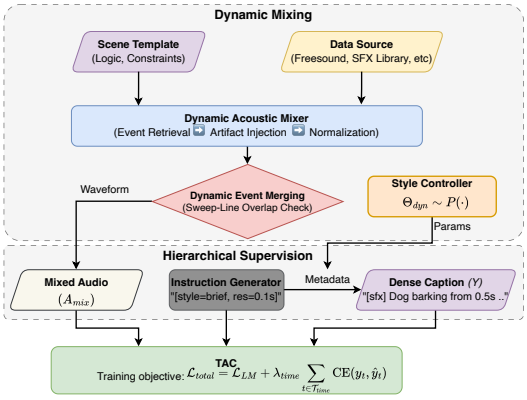


Figure 2: **The TAC Training Pipeline.** Stage 1 synthesizes complex audio mixtures via our *Dynamic Acoustic Mixer*. In Stage 2, a *Style Controller* stochastically samples "description styles" (Keyword vs. Brief vs. Detailed) and timing resolutions, generating a diverse curriculum of instruction-tuned prompts.

template might require that the music source plays continuously, a speech source can occur randomly throughout (while never overlapping with another speech stream), and the sound effects source is restricted to background ambience, keyboard clicking, phone ringing. While the actual underlying sources are random, by tuning these templates we can make an endless combination of targeted synthetic mixtures for specific tasks. Our mixer allows for flexible control of various properties, such as number of concurrent sounding events, amount of reverberation and other signal-level augmentation, and number of repeats of an event.

Finally, precise temporal grounding is achieved via RMS-based activity detection with an activity threshold of δ_{act} (a proxy for loudness) unlike metadata which is often used in literature and relevant works. For every instantiated event e_i, we compute a continuous activity map $M_i(t)$. Given a merge threshold $\delta_{\text{merge}} \sim \mathcal{U}(0.1,\ 1.0)$, in seconds, if two activations of the same event are separated by a gap $g < \delta_{\text{merge}}$, they are fused into a single timestamped segment. While one can choose δ_{act} and δ_{merge} statically before training, we instead choose them per example during training, and condition the model on the chosen values.

In practice, we generate $\sim$150K synthetic mixtures ($\sim$1.2K hours) for training, with an average of 5 events per clip and $\sim$20s average clip duration. The full mixing procedure is described in Algorithm 1, and a comparison with existing temporally grounded audio-caption resources is provided in Table 9 (both in Appendix E).

3.2 Multitask prompts and output format

Instead of fixing the tasks statically at the beginning of the training (for example deciding that model must detect sounds with a granularity of 0.25s), we instead sample from a set of multitask prompts, and modify the target caption accordingly. There are 4 high-level properties for each task that we can control per training sample:

1. Style: we sample from various caption styles for each event in the soundscape. These styles can be brief ("Dog barks"), keywords ("Dog"), or detailed ("A dog barks aggressively twice").

2. Merge threshold: δ_{merge} dictates how close an events offset must be near its closest onset before they are merged into one item. For example, this can decide if two quick utterances are detected as one event (e.g. "Speech from 5.0s to 10.0s", or two events (e.g. "Speech from 5.0s to 7.0s, 8.0s to 10.0s").

3. Activity threshold: δ_{act} controls how quiet a sound must get to its minimum before it is considered "off". This has an effect on sounds that are intermittent, but do not go all the way to silence, such as explosions, whooshes, or other sound design elements. A high activity threshold will break up sounds into many events; a low activity threshold will keep them as one.

4. Time resolution: We round off start and end times randomly when deciding what ground truth is. For example, we can round off to the nearest half second, or tenth of a second. This controls the resolution at which we want to caption the audio.

Exemplar Training Pair

Input Audio (Synthetic Mixture):
[0.0-10.0s] Background Jazz Music + [0.5-2.3s] Dog Bark + [5.0-5.5s] Car Horn

Conditioning Prompt:

```
<|im_start|>user Describe all
events in the audio.  Give start
and end times.  [style=brief,
merge=0.25s, activity=0.05,
resolution=0.10s] <|im_end|>
```

Target Output (Ground Truth):

```
<|im_start|>assistant 3 events
total.  2 events overlap.  2
sound effects, 1 music.  [music]
Background jazz music plays softly
from <|0.00|>s to <|10.00|>s.
[sfx] A dog barks aggressively
from <|0.50|>s to <|2.30|>s.  [sfx]
A car horn honks briefly from
<|5.00|>s to <|5.50|>s.<|im_end|>
```

Figure 3: An example of a synthetically generated training pair. Note how the "Reasoning Header" ("3 events total...") is algorithmically derived from the composition metadata, teaching the model to summarize before detailing.

As shown in Algorithm 1, during training, we randomly sample a *Caption Style* $\mathcal{S} \in \{\text{KEYWORDS}, \text{BRIEF}, \text{DETAILED}\}$ and a set of *Timing Parameters* (resolution δ_{res}, merge threshold δ_{merge}, and activity threshold δ_{act}). The instruction prompt P is conditioned on these parameters (e.g., `"[style=brief, resolution=0.1s]"`). This instruction tuning allows us to control the model's output density at inference time. We form the target sequence in a token efficient way by concatenating all start and end times for each event as a comma separated list with the description. Captions are ordered by start time. Each caption is associated

with a "type" (music, sfx, speech, background), which is prepended to the caption as '[type]'. An example of an input/output pair can be seen in Figure 3. The structured output of TAC can be easily parsed into a data structure, and used reliably for downstream tasks.

3.3 TAC Architecture and Training

Though any backbone can be used, we use Qwen2-Audio [Chu et al., 2023] for ours, freezing the base model and fine-tuning via Low-Rank Adaptation (LoRA) [Hu et al., 2022] on linear layers. Standard LALMs, including our backbone Qwen2-Audio, are trained on broad in-the-wild data. While effective for general audio, they miss fine-grained, domain-specific acoustics (e.g., distinguishing an "industrial hum" from a "sci-fi drone"), undermining dense captioning. Therefore, we decided to continue pretraining on Qwen2Audio on a large corpus of high-fidelity licensed single-source audio (e.g. an explosion sound effect, or a music track) paired with descriptive captions at varying levels of detail. We generated these captions from metadata, following the approach laid out in AudioCards Sridhar et al. [2026]. We expanded these captions to an instruction tuning set using off-the-shelf LLMs (GPT-OSS-120B Agarwal et al. [2025] and Qwen-32B-VL Bai et al. [2025]) with a variety of questions, such as identification ("What is the source of this sound?"), and description ("Describe the mood.").

Standard cross-entropy loss is often insufficient for dense captioning, as it treats short-duration timestamp tokens equally with semantic tokens. To strictly enforce temporal precision, we tokenize timestamps as atomic special tokens (e.g., <|1.23|>), as done in prior work Radford et al. [2023], Chu et al. [2023]. Unlike prior work, we introduce a weighted loss objective $\mathcal{L}_{total}$:

$$\mathcal{L}_{total} = \mathcal{L}_{LM} + \lambda_{time} \sum_{t \in \mathcal{T}_{time}} \text{CE}(y_t, \hat{y}_t) + \lambda_{spk} \sum_{s \in \mathcal{T}_{spk}} \text{CE}(y_s, \hat{y}_s) \quad (1)$$

where $\mathcal{T}_{time}$ represents the set of indices corresponding to timestamp tokens, $\mathcal{T}_{spk}$ the indices of speaker identity tokens (e.g., "Speaker 1", "Speaker 2"), λ_{time} upweights temporal alignment errors, and λ_{spk} upweights speaker identity token errors during speaker diarization fine-tuning (λ_{spk}=0 for captioning/SED, λ_{spk}=10.0 for diarization; see Section 4.2). Finally, while TAC can be directly trained for speech transcription, we transcribe detected speech separately. Each TAC-detected speech interval is cropped and passed independently to Whisper Radford et al. [2023]; TAC supplies timestamps and speaker attribution, while Whisper supplies only text, so no cross-model timestamp alignment is required. For overlap, TAC emits separate events and crops per speaker, supervised by the Overlapping Speakers template. If voices cannot be separated acoustically, transcription degrades as in other ASR systems and we do not claim reliable per-speaker transcripts, although event boundaries and attribution still come from TAC. For example, "Male voice whispering from 1.0s to 8.0s" expands to "Male voice whispering from 1.0s to 8.0s <speech>Do you want to know a secret?</speech>".

3.4 TAC-V: TAC with Visuals

To demonstrate the extensibility of TAC, we introduce **TAC-V**, a pipeline that fuses the high temporal-precision outputs of TAC with a Visual Language Model (VLM) for temporally dense audio-visual captions. The pipeline processes audiovisual inputs to produce timestamped, visually-grounded captions via five distinct stages. We first extract the audio and sample video frames at a configurable frame rate (we choose 2fps). For video resolution, we alternate between 360p and 240p for every other frame, to stay within the effective token limit of our chosen VLM.

Audio captioning: We process the audio by chunking it into $20s$ non-overlapping chunks. Each chunk is processed in parallel with TAC. Unlike other audio LMs, TAC provides precise timestamped events tagged by category (e.g.,

TAC-V: Fused Audio-Visual Captions

```
1. [0.2s - 2.0s] [music] Epic
orchestral surge as metallic
studio logo glows (0.99)
2. [3.8s - 6.2s] [sfx]
Explosion erupts as debris
flies through smoke (0.83)
3. [7.4s - 9.0s] [speech]
Woman (whispering, on-screen)
with furrowed brow (0.89)
   <speech>"Their cruelty is
all I've known."</speech>
```

Figure 4: An example output from our cascaded Audio-Visual pipeline. Note the integration of visual details ("metallic studio logo", "furrowed brow") with precise audio events, and the inclusion of FLAM confidence scores (e.g., 0.99) alongside aligned transcriptions.

[speech]). We augment the output of TAC by transcribing all detected speech events. Finally, we score every event by using FLAM [Wu et al., 2025b], which assigns a confidence score $c \in [0, 1]$ to each detected event. This serves as a signal for the downstream VLM: low confidence scores flag ambiguous events that require visual verification.

Audio-driven video captioning: From TAC, we create a "shot-list" of audio events, ordered by time, with precise timestamps, types, captions, and transcriptions. We augment this shot-list with visual shot boundaries (points where the scene changes in the video), placing them in the scene as visual markers. This ensures even coverage across an entire video, and aids the model in distinguishing continuous audio events from changing visual perspectives. We feed the video frames, the timestamped shot-list and confidence scores into Qwen3-VL-32B. Using a specialized *Chain-of-Thought* prompt, the VLM performs *Hallucination Correction* (using visuals to resolve acoustic ambiguity) and *Visual Grounding* (linking sounds to visible sources). Figure 4 illustrates the final structured output; the complete prompt and worked example appear in Appendix H.3. The fused captions combine acoustic classification (e.g., [sfx]), visual grounding (e.g., "debris flies"), and speech transcription into a unified timeline.

3.5 Evaluation

Evaluating dense audio captioning is challenging because a single acoustic scene can be validly described at multiple levels of granularity, making standard metrics such as CIDEr Vedantam et al. [2015], SPICE Anderson et al. [2016], and SPIDEr Liu et al. [2017] insufficient for capturing temporal precision or factual correctness. To address this limitation, we evaluate TAC along three axes: *semantic alignment*, *temporal precision*, and *robustness*.

Semantic alignment: Exact string matching is insufficient for dense captions Kumar et al. [2025] (e.g., "car engine" vs. "vehicle idling" should be a match). We propose a reference-based metric using an LLM as a judge. For every predicted event e_{pred} and ground truth event e_{gt}, we compute a *Semantic Similarity Score* $S_{sem} \in [0, 1]$:

$$S_{sem}(e_{pred}, e_{gt}) = \text{LLM}_{\text{judge}}(d_{pred}, d_{gt}) \tag{2}$$

We use GPT-OSS-120B Agarwal et al. [2025] at temperature zero as the judge. It scores each description pair from 0 to 10 using a four-band rubric with worked examples, and we normalize the result to $[0, 1]$. Matching is one-to-one and greedy: a candidate pair must have semantic similarity of at least 0.5 and either temporal overlap or containment; eligible candidates are ordered by semantic similarity weighted by temporal overlap. Thus semantic agreement alone cannot produce a match. We additionally report sentence-embedding-based FENSE, alongside METEOR, CIDEr, and SPICE, to complement the compositional LLM judgment. **Temporal precision:** To rigorously test the model's ability to localize events, we adapt Sound Event Detection (SED) metrics [Mesaros et al., 2016, Temko et al., 2006]. After semantic alignment with a ground truth reference caption, we treat the generated captions as detection outputs and compute:

(i) Segment-Based F1 (SegF1): Evaluates activity detection at a 100ms resolution. This measures how well the predicted duration matches the ground truth, regardless of the exact start/end times.

(ii) Event-Based F1 (EvtF1): Treats each caption segment as a discrete event. A prediction is counted as a True Positive (TP) only if its onset is within a ± 1.0s window (or *collar*) of the ground truth onset.

Robustness & Hallucination: Hallucination remains a major challenge for LALMs Chen et al. [2025b]. These models frequently produce temporally misaligned descriptions, invent subtle sound effects, misinterpret overlapping events, or confuse acoustically similar sources. To assess performance in the absence of ground truth (or to detect hallucinations where the ground truth is silent), we utilize FLAM [Wu et al., 2025b] for reference-free evaluation. Reference matching remains primary for EvtF1, SegF1, and semantic similarity, but incomplete annotations can mark an unannotated-yet-present sound as a false positive and cannot distinguish it from an acoustically absent invention; FLAM complements it by scoring each predicted event directly against the audio. We define the *Hallucination Rate (Hal%)* as the percentage of predicted events where the FLAM confidence score drops below $\tau = 0.25$, fixed before evaluation. A sweep from 0.15 to 0.50 confirms that the ranking is stable from $\tau = 0.20$ upward (Appendix B). We report **confidence (conf)** - the maximum audio-text similarity within the predicted time range - and **specificity (spec)** - the *minimum* similarity across the predicted range. A high specificity indicates the model is not just detecting a peak, but accurately

	Multitask	Pretrained	Templates	Acoustic Sim	TACOS	Diarization	Iters	LoRA	TS Wt	**EvtF1↑**	SegF1	**Hal%↓**	Conf	Spec
Ours (TAC)	✓	✓	✓	✓	✓	✗	5k	128	5.0	**.50**	**.71**	4.9	0.89	0.74
Ours (TAC-Dia.)	✓	✓	✓	✓	✓	✓	5k	128	5.0	.49	**.71**	3.0	0.91	0.77
Ablations														
✗ Multitask	✗	✓	✓	✓	✓	✗	5k	128	5.0	.45	.72	7.0	0.87	0.70
(merge=0.1)	✗	✓	✓	✓	✓	✗	5k	128	5.0	.41	.71	13.8	0.80	0.70
✗ Pretrained	✓	✗	✓	✓	✓	✗	5k	128	5.0	.49	.70	8.8	0.85	0.70
✗ Templates	✓	✓	✗	✓	✓	✗	5k	128	5.0	.47	.71	2.2	0.93	0.78
✗ Acoustic Sim	✓	✓	✓	✗	✓	✗	5k	128	5.0	.49	.71	5.3	0.89	0.75
✗ TACOS	✓	✓	✓	✓	✗	✗	5k	128	5.0	.42	.68	7.6	0.85	0.70
LoRA Rank	✓	✓	✓	✓	✓	✗	5k	256	5.0	.48	.70	3.5	0.90	0.75
						✗	5k	64	5.0	.49	.71	4.8	0.89	0.74
						✗	5k	8	5.0	.19	.66	36.0	0.58	0.54
Timestamp weight	✓	✓	✓	✓	✓	✗	5k	128	1.0	.48	.71	4.2	0.91	0.76
						✗	5k	128	10.0	.48	.71	5.8	0.88	0.73
Iterations	✓	✓	✓	✓	✓	✗	10k	128	5.0	.47	.70	5.2	0.89	0.75
						✗	2.5k	128	5.0	.46	.70	8.0	0.85	0.72
Baselines														
Gemini 3 Pro	–	–	–	–	–	–	–	–	–	.42	.64	6.1	0.84	0.66
Qwen3-Omni	–	–	–	–	–	–	–	–	–	.37	.66	7.3	0.84	0.62
Audio Flamingo 3	–	–	–	–	–	–	–	–	–	.27	.55	11.6	0.73	0.59

(a) Training Ablations & Baselines

Style	Merge	Activity	Resolution	**EvtF1↑**	SegF1	**Hal%↓**	Conf	Spec
brief	**0.25**	**0.05**	**0.10**	**.50**	**.71**	**4.9**	**0.89**	**0.74**
detailed	0.25	0.05	0.10	.49	.71	8.0	0.86	0.72
keywords	0.25	0.05	0.10	.47	.66	1.3	0.89	0.78
brief	0.10	0.05	0.10	.31	.66	20.2	0.73	0.67
brief	0.50	0.05	0.10	.48	.72	4.0	0.90	0.74
brief	1.00	0.05	0.10	.42	.72	4.7	0.89	0.69
brief	0.25	0.01	0.10	.49	.72	4.7	0.89	0.74
brief	0.25	0.10	0.10	.49	.70	5.5	0.88	0.76
brief	0.25	0.20	0.10	.45	.70	4.5	0.90	0.76
brief	0.25	0.05	0.01	.43	.71	11.8	0.83	0.73
brief	0.25	0.05	0.50	.48	.70	5.4	0.88	0.77

(b) Inference Parameter Sweeps

Table 1: **Comprehensive Evaluation.** (a) Training ablations and baselines. (b) Post-hoc controllability analysis on the TACOS test set; the bold default was fixed on a separate development set before evaluation and was not selected using this sweep.

describing the entire duration of the event. On 100 human-rated TACOS samples, the LLM semantic judge correlates with human ratings at Spearman 0.81/Kendall 0.76, while FLAM confidence reaches $0.73/0.69$ (Appendix B).

4 Experiments

Training Setup. We train TAC on a cluster of 8 NVIDIA A100 (80GB) GPUs, with a global effective batch size of 32. We freeze the pre-trained backbone and only fine-tune low-rank adapters (LoRA) with a rank of r=128 and alpha α=256. Optimization is performed using AdamW with a peak learning rate of 5e-5, following a cosine decay schedule with 1000 steps of linear warmup. The training data is a multi-stream mixture comprising our synthetic audio mixtures, SED data, TACOS, and real-world diarization data ($\sim$943 hours from 10 corpora across 6 languages; see Appendix E.2). We set $\lambda_{time} = 5.0$ and $\lambda_{spk} = 10.0$. We ensured all experiments started from the exact same seed, with identical data.

Evaluation Datasets. To comprehensively assess the diverse capabilities of TAC, we employ a multi-faceted suite of evaluation benchmarks. We evaluate *timestamped dense captioning* performance using the test set from TACOS Primus et al. [2025]. To assess our TAC$\rightarrow$LLM cascade, we leverage audio understanding & reasoning benchmarks including MMAU Sakshi et al. [2025], MMAR Ma et al. [2025a], MMSU Wang et al. [2025], and MMAU-Pro Kumar et al. [2025]. We evaluate our TAC-V$\rightarrow$LLM cascade (Section 3.4) on Daily-Omni Zhou et al. [2025], World-Sense Hong et al. [2026], Video-Holmes Cheng et al. [2025a], and AVHBench Sung-Bin et al. [2025]. For TACOS Primus et al. [2025], we adopt the evaluation metrics described in Section 3.5, while for all other benchmarks we adopt their standard metrics.

4.1 Dense Captioning

We evaluate TAC on the held-out test set of the **TACOS** benchmark, and zero-shot performance on AudioCaps Kim et al. [2019] and Clotho-v2 Drossos et al. [2020], unlike baselines trained on these datasets (Table 3a), using METEOR, CIDEr, SPICE, and FENSE Zhou et al. [2022]. For TACOS, every baseline was evaluated both zero-shot and few-shot with in-context examples of the required output format, and we report whichever condition scored higher. This permissive selection was made on the test set for baselines, whereas TAC's configuration was fixed on a separate development set; the comparison is therefore optimistic toward the baselines. Timestamped captioning results are summarized in Table 1.

Dataset	Model	METEOR	CIDEr	SPICE	FENSE
Clotho-v2	Audio Flamingo 3	13.74	43.94	11.28	52.1
	Qwen3-Omni-Captioner	12.13	43.85	10.46	52.43
	Audio Flamingo Next	18.58	46.45	13.84	56.21
	TAC (ours, zero-shot)	**19.07**	**48.83**	**14.07**	**58.4**
AudioCaps	Audio Flamingo 3	23.28	70.04	16.56	58.2
	Qwen3-Omni-Captioner	19.46	65.84	17.16	59.27
	Audio Flamingo Next	24.39	**71.16**	16.92	62.4
	TAC (ours, zero-shot)	**24.53**	68.29	**17.87**	**63.1**

(a) Standard Audio Captioning

Model	VDC Detailed		V-SAL 2 Test		Omni-Cloze
	%Acc↑	Score↑	%Miss↓	%Hall↓	Total↑
GPT-4o	46.3	2.5	17.0	14.2	–
Gemini 1.5 Pro	43.1	2.2	21.8	16.5	–
Qwen2.5-Omni-7B	39.7	2.2	26.3	21.7	–
Video-LLaMA 2	29.7	1.4	56.8	8.9	6.6
Video-SALMONN 2	46.1	2.5	10.0	12.9	3.3
Omni-Captioner	55.0	2.7	17.8	10.9	56.4
Gemini 3.1 Pro	56.4	3.5	12.2	7.4	–
TAC-V (ours)	**59.8**	**3.6**	**8.7**	**5.2**	**61.9**

(b) Dense Audiovisual Captioning

Table 3: **Captioning Evaluation.** (a) Standard audio captioning on Clotho-v2 and AudioCaps. (b) Dense audiovisual captioning on VDC Chai et al. [2024], Video-SALMONN 2 Test Set Tang et al. [2025], and Omni-Cloze Ma et al. [2025b] (Total = avg. of Visual, Audio, and AV subsets).

Comparison with State-of-the-Art: We first analyze the bottom section of Table 1. TAC achieves a new state-of-the-art across all major temporal and semantic metrics, significantly outperforming previous state-of-the art models. The most striking improvement is in temporal grounding. We observe that for Event F1 score (EvtF1) our TAC beats Qwen 3 Omni by 0.14 F1 Score, and Gemini 3 Pro by 0.08 F1 Score. Outside of temporal grounding, TAC also out-performs in text-audio similarity (0.89 vs 0.84), and Segment F1 score (0.71 vs 0.66 for Gemini 3 Pro / 0.64 for Qwen3-Omni). Competing models perform decently at "global" recognition, but fail to localize events precisely in dense mixtures. Our approach yields the lowest Hallucination Rate (**4.9%**), nearly half that of the open-source baseline Audio Flamingo 3 (11.6%) and significantly lower than Gemini 3 Pro (6.1%). Furthermore, our high Specificity score (0.74) indicates that TAC does not merely "spot" keywords but accurately describes the full duration of acoustic events.

Ablation study: We conduct a thorough ablation study of TAC, varying each component one by one and studying its impact on temporal grounding and semantic similarity as shown in Table 1a. We observe that multitask prompts are critical: static tasks drop EvtF1 from 0.50 to 0.45, and a poor merge threshold (0.1s) causes severe degradation (0.41 EvtF1, 13.8% Hal). Pretraining cuts hallucination in half ($8.8\%\rightarrow4.9\%$), while scene templates improve EvtF1 ($0.47\rightarrow0.50$) at the cost of slightly higher hallucination - without templates, the model becomes overly conservative. LoRA rank 128 seemed to work best and rank 8 caused collapse. Training beyond 5k iterations also degraded performance due to overfitting. The timestamp loss weight λ_{time}=5.0 balances EvtF1 and hallucination, and removing TACOS causes a large drop (0.42 EvtF1), confirming that real-world annotations helps generalization.

Prompt controllability: Table 1b is a post-hoc analysis of how user controls affect the fixed TAC checkpoint, not a test-set hyperparameter search. The default (brief, merge 0.25s, activity 0.05, resolution 0.10s) was selected on a separate development set before evaluation and is used throughout. Moving merge to 0.1s sharply lowers Event F1 and raises hallucination; keyword captions have low hallucination because they are conservative; and a higher activity threshold widens event spans, trading temporal recall for confidence.

4.2 Speaker Diarization

Because TAC produces timestamped speech events with precise boundaries, it can be directly fine-tuned for speaker diarization - the task of determining "who spoke when." We fine-tune on a multi-corpus mixture of ∼943 raw hours (∼468 effective) from 10 corpora across 6 languages with per-family equalization (see Appendix E.2 for full details). At inference, TAC-Dia. first counts speakers, then diarizes conditioned on the predicted count. We evaluate on four benchmarks using Diarization Error Rate (DER%) with a strict collar of 0.25s. Table 2 shows that TAC achieves the lowest pooled DER across all four datasets, outperforming dedicated systems including pyannote-3.1 Plaquet and Bredin [2023], DiariZen [Raghav, 2026], and Sortformer [Medennikov, 2025]. The count-then-diarize strategy consistently improves both DER and speaker-counting accuracy (K-acc) over direct diarization (see Appendix H.4 for the prompt and sample outputs).

Model	VoxConv.	AMI	AISH.-4	Bazinga	Avg.	K-acc%
TAC-Dia. (ours)	6.33	11.10	6.81	**28.29**	**10.16**	71.0
w/o sp. count	6.37	**10.74**	6.92	29.75	10.37	69.7
DiariZen	**5.91**	22.51	**5.70**	38.98	12.60	68.5
Sortformer-strm.	6.15	24.96	8.45	46.39	14.57	**71.9**
Pyannote c-1	7.55	12.83	6.88	47.30	13.84	61.2
Pyannote-3.0	7.47	13.48	7.06	48.00	14.00	59.6
Pyannote-3.1	7.44	13.43	6.94	48.29	14.00	59.8
Reverb-v2	16.33	36.25	13.74	150.96	37.80	54.0

Table 2: **Speaker Diarization** (DER%↓, strict collar=0.25s). K-acc = % windows with exactly correct speaker count. TAC first counts speakers, then diarizes, achieving the lowest pooled DER across all four datasets.

Benchmark	Prior Methods		TAC + Text-only Reasoner	
	Model	Score	+ Qwen3	+ Gemini3
MMAU	Audio Thinker	**75.9**	73.9	72.2
Sound		78.8	**79.7**	79.6
Music		**73.8**	62.6	63.4
Speech		75.2	**79.3**	73.6
MMAU-mini	Gem. Flash→Gem. 3 Pro	68.5	71.7	**72.2**
	Gemini 2.5 Flash	64.3		
MMAR	Gem. Flash→Gem. 3 Pro	65.9	60.1	**71.9**
	Audio Flamingo 3	60.1		
MMSU	Audio Flamingo 3	62.3	65.0	**72.4**
MMAU-Pro	Gemini 2.5 Flash	59.2	62.5	**62.9**

(a) Audio Understanding & Reasoning

Benchmark	Prior Methods		Describer + Text-only Reasoner				
	Model	Score	VLM + Qwen3	V-LLaMA2 + Gem3	V-SAL2 + Gem3	TAC-V + Qwen3	TAC-V + Gem3
Daily-Omni	Qwen3-Omni	76.2	51.5	41.4	68.4	72.9	**77.9**
	Gemini 2.5 Flash	72.7					
	OmniVinci	66.5					
World-Sense	Gemini 2.5 Pro	**65.1**	37.4	40.7	42.2	45.7	58.6
	OmniVinci	48.2					
Video-Holmes	Qwen3-Omni	57.3	45.6	39.3	47.6	47.7	**59.2**
AVHBench (AVH)	PandaGPT	58.5	70.8	70.4	71.2	79.8	**81.7**
AVHBench (VAH)	PandaGPT	61.3	51.8	46.2	41.4	76.1	**76.6**
AVHBench (AVM)	OneLLM	60.1	50.5	53.1	55.7	56.7	**61.6**
AVHBench (AVC)	Video-LLaMA 2	14.0	12.9	7.4	6.9	**22.6**	20.6

(b) Audio-visual Understanding & Reasoning

Table 4: **Downstream Reasoning Benchmarks.** We compare prior SOTA end-to-end and cascade methods against our cascade approach: TAC/TAC-V captions fed to a text-only reasoner. We include cascade baselines (e.g., Gemini Flash→Gemini 3 Pro, Video-LLaMA 2/Video-SALMONN 2→Gemini 3 Pro) for fair comparison using the same downstream reasoner.

5 Describe-Then-Reason

We now turn to using TAC and its audiovisual extension TAC-V as a semantic bridge to a text-only reasoner. Here, TAC(-V) converts audio or video into a precise timestamped text representation; the reasoner never sees the original media. We call this paradigm "describe-then-reason". We evaluate Qwen3-8B, Qwen3-Next-80B-A3B-Thinking Yang et al. [2025], and Gemini 3 Pro Team and Google [2025]. Scaling helps while reasoning is the bottleneck, but can saturate when the caption omits task-specific evidence. The small-reasoner and matched-captioner controls in Appendix B separate caption quality from reasoner capacity.

5.1 Audio Understanding & Reasoning

Table 4a shows results on complex audio understanding and reasoning benchmarks. With Qwen3-Next-80B, TAC reaches 73.9% on MMAU; Gemini 3 Pro yields 71.9% on MMAR, 72.4% on MMSU, and 62.9% on MMAU-Pro. A much smaller Qwen3-8B still reaches $64.2/51.2/61.3$ on MMAU/MMAR/MMSU. A dedicated-captioner control, Qwen3-Omni-Captioner→Gemini 3.1 Pro, obtains $65.4/67.9/57.1$ on MMAU-mini/MMAU/MMAR, below TAC captions with Gemini 3 Pro ($72.2/72.2/71.9$). Thus both caption quality and reasoner capacity matter (Appendix B).

5.2 Audiovisual Understanding & Reasoning

We apply TAC-V (Sec. 3.4) to obtain dense timestamped audiovisual captions. We evaluate the quality of our generated audiovisual captions by using them as the *sole input* for downstream reasoning tasks. In this setup, the reasoning Large Language Model (LLM) *sees no video or audio*; it must answer complex questions based entirely on the text description generated by TAC-V. Table 4b presents results against SOTA native multimodal models. Our text-based cascade achieves SOTA on *Daily-Omni* and *Video-Holmes*, and significant gains on *AVHBench*, which explicitly measures cross-modal hallucination. By separating event detection from visual grounding, our pipeline regularizes against the hallucinations common in end-to-end models (see Section G for examples). Crucially, alternative cascades using the same Gemini 3 Pro reasoner but weaker captioners all underperform TAC-V, confirming that the gains stem from our caption quality (Table 4b). We additionally evaluate on dedicated dense captioning benchmarks (Table 3b), where TAC-V surpasses the prior SOTA Omni-Captioner Ma et al. [2025b] on VDC Detailed Chai et al. [2024] and Omni-Cloze, and achieves the lowest miss and hallucination rates on the Video-SALMONN 2 Test Set Tang et al. [2025].

6 Conclusion, Limitations, and Future Work

In this work, we introduced TAC, a model that bridges the gap between raw acoustic signals and high-level reasoning through temporal dense captioning. We showed that robust temporal grounding can be learned from purely synthetic mixtures. We further extended TAC with a VLM, producing TAC-V, which generates rich, high-quality dense audio-visual captions. TAC achieves state-of-the-art performance on dense captioning benchmarks–surpassing frontier systems such as Gemini 3 Pro. When cascaded with text-only LLMs, both TAC and TAC-V serve as powerful semantic bridges for

downstream reasoning, unlocking expert-level state-of-the-art performance on audio and audio-visual reasoning benchmarks, respectively.

Despite these advancements, synthetic training can cause domain shift: the model sometimes over-predicts dramatic events (e.g., gunshots) in mundane videos and lacks fine-grained musical precision (e.g., chord progressions). The mixer partly targets this risk with convolution over 20.6K measured room impulse responses, sparse room-tone scenes, and templates spanning high-noise, short-segment, and event-density extremes. All evaluations use real recordings; zero-shot AudioCaps/Clotho and the held-out MSDWild results provide encouraging evidence, but do not cover all environments. Future work will study unsupervised domain adaptation and calibration of event priors. Describe-then-reason is also reusable: one caption can answer many questions without reprocessing the media, though performance saturates when the caption omits the evidence a stronger reasoner needs.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *CoRR*, abs/2508.10925, 2025.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Proc. ECCV*, pages 382–398, 2016.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *CoRR*, abs/2502.13923, 2025.

Wenhao Chai et al. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. In *Proc. ICML*, 2023.

William Chen et al. Audiochat: Unified audio storytelling, editing, and understanding with transfusion forcing. *arXiv preprint arXiv:2602.17097*, 2026.

Xinlong Chen et al. Avocado: An audiovisual video captioner driven by temporal orchestration. In *The Fourteenth International Conference on Learning Representations*, 2025a.

Yanxi Chen, Wenhui Zhu, Xiwen Chen, Zhipeng Wang, Xin Li, Peijie Qiu, Hao Wang, Xuanzhao Dong, Yujian Xiong, Anderson Schneider, et al. Aha: Aligning large audio-language models for reasoning hallucinations via counterfactual hard negatives. *CoRR*, abs/2512.24052, 2025b.

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

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. In *Proc. NeurIPS*, 2025b.

Zesen Cheng et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *CoRR*, abs/2311.07919, 2023.

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. *CoRR*, abs/2407.10759, 2024.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *CoRR*, abs/2504.18425, 2025.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *Proc. ICASSP*, 2020.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proc. EMNLP*, 2024.

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

Sreyan Ghosh, Sonal Kumar, Chandra Kiran Reddy Evuru, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Reclap: Improving zero shot audio classification by describing sounds. In *Proc. ICASSP*, 2025b.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang-gil Lee, Jaehyeon Kim, Aya Aljafari, et al. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*, 2026.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *Proc. NeurIPS*, 2025.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. In *Proc. Interspeech*, 2021.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. Listen, think, and understand. In *Proc. ICLR*, 2024.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proc. CVPR*, 2024.

Shawn Hershey, Daniel P W Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. The benefit of temporally-strong labels in audio event classification. In *Proc. ICASSP*, 2021.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *Proc. ICLR*, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proc. ICLR*, 2022.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proc. CVPR*, 2016.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proc. NAACL*, 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. In *Proc. ICML*, 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 *Proc. ICASSP*, 2025.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, et al. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *CoRR*, abs/2508.13992, 2025.

Kai Li, Jintao Cheng, Chang Zeng, Zijun Yan, Helin Wang, Zixiong Su, Bo Zheng, and Xiaolin Hu. A semantically consistent dataset for data-efficient query-based universal sound separation. *arXiv preprint arXiv:2601.22599*, 2026.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of SPIDEr. In *Proc. ICCV*, 2017.

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. In *Proc. ICML*, 2025a.

Ziyang Ma et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. *arXiv preprint arXiv:2510.12720*, 2025b.

Ivan others Medennikov. Streaming sortformer: Speaker cache-based online speaker diarization with arrival-time ordering. *arXiv preprint arXiv:2507.18446*, 2025.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2024.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. *Applied Sciences*, 2016.

Alexis Plaquet and Hervé Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. *arXiv preprint arXiv:2310.13025*, 2023.

Paul Primus, Florian Schmid, and Gerhard Widmer. Tacos: Temporally-aligned audio captions for language-audio pretraining, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 28492–28518. PMLR, 2023.

Nikhil Raghav. Diarizen explained: A tutorial for the open source state-of-the-art speaker diarization pipeline. *arXiv preprint arXiv:2604.21507*, 2026.

S 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 *Proc. ICLR*, 2025.

Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello. Scaper: A library for soundscape synthesis and augmentation. In *Proc. WASPAA*, 2017.

Sripathi Sridhar, Prem Seetharaman, Oriol Nieto, Mark Cartwright, and Justin Salamon. Audiocards: Structured metadata improves audio language models for sound design. In *Proc. ICASSP*, Barcelona, Spain, 2026.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *CoRR*, abs/2305.16355, 2023.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *Proc. ICLR*, 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 *Proc. ICLR*, 2024.

Changli Tang et al. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Gemini Team and Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

Andrey Temko, Robert Malkin, Christian Zieger, Dušan Macho, Climent Nadeu, and Maurizio Omologo. Clear evaluation of acoustic event detection and classification systems. In *Proc. International Evaluation Workshop on Classification of Events, Activities and Relationships*, 2006.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proc. CVPR*, pages 4566–4575, 2015.

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

Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. Audio-thinker: Guiding audio language model when and how to think via reinforcement learning. *CoRR*, abs/2508.08039, 2025a.

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. In *Proc. ICML*, 2025b.

Jin Xu, Zhifang Guo, Hangrui Hu, et al. Qwen3-omni technical report. *CoRR*, abs/2509.17765, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm, 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. *CoRR*, abs/2512.23808, 2025.

Zelin Zhou et al. Can audio captions be evaluated with image caption metrics? In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 981–985. IEEE, 2022.

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

A Appendix

- Section C: Impact Statement
- Section D: Baselines
- Section E: Dataset & Pipeline Details
- Section F: Qualitative Analysis: Audio Understanding & Reasoning
- Section G: Qualitative Analysis: Audio-Visual Understanding
- Section H: Prompts
- Section I: LLM Usage
- Section B: Additional Validation and Efficiency Results

B Additional Validation and Efficiency Results

Human validation and threshold sensitivity. For 100 sampled TACOS predictions, human raters judged whether every predicted event was present in the audio. Both automatic measures correlate strongly with these judgments: Spearman/Kendall correlations are $0.81/0.76$ for the LLM semantic judge and $0.73/0.69$ for FLAM confidence. Table 5 varies the FLAM threshold without changing predictions. TAC has the lowest Hal% from $\tau = 0.20$ upward. At $\tau = 0.15$, TAC and Gemini 3 Pro are effectively tied; this rate should be read with recall because Gemini is more conservative (EvtF1/SegF1 $0.42/0.64$ versus TAC's $0.50/0.71$).

Table 5: Hallucination rate (Hal%↓) across FLAM thresholds on TACOS.

τ	.15	.20	.25	.30	.40	.50
TAC (ours)	3.8	**4.3**	**4.9**	**5.5**	**7.0**	**8.3**
Gemini 3 Pro	**3.7**	5.1	6.1	7.2	9.6	12.3
Audio Flamingo 3	8.9	10.2	11.6	13.1	16.7	21.6

Cross-corpus diarization. We evaluate zero-shot on MSDWild, which is absent from the training mixture, using 30-second windows. TAC-Dia improves DER on both splits: $17.80/42.10$ (few/many speakers), compared with pyannote's $21.96/43.15$ and DiariZen's $22.42/45.70$.

Reasoner and captioner controls. Table 6 reports the discussion-period controls. Holding TAC captions fixed, moving from Qwen3-8B to Qwen3-Next-80B substantially improves all three benchmarks; Gemini 3 Pro further improves reasoning-heavy MMAR and MMSU but is flat on MMAU. Conversely, even with the stronger Gemini 3.1 Pro reasoner, the untimestamped Qwen3-Omni-Captioner cascade trails TAC→Gemini 3 Pro.

Table 6: Describe-then-reason controls (accuracy, %). A dash denotes an unreported result.

Describer → reasoner	MMAU	MMAU-mini	MMAR	MMSU
Qwen3-Omni-Captioner → Gemini 3.1 Pro	67.9	65.4	57.1	–
TAC → Qwen3-8B	64.2	65.1	51.2	61.3
TAC → Qwen3-Next-80B	**73.9**	71.7	60.1	65.0
TAC → Gemini 3 Pro	72.2	**72.2**	**71.9**	**72.4**

Latency. Table 7 reports wall-clock seconds on one A100. TAC is served with vLLM; the VLM captions 30-second chunks at 2 fps; Whisper large-v3-turbo processes only detected speech; API network time is included for the reasoner. TAC accounts for under 10% of the 10-minute end-to-end total. The caption is computed once and can be reused for multiple questions.

Table 7: Pipeline latency in seconds.

Input	TAC	Whisper	VLM (TAC-V)	Gemini 3.1 Pro
30 s	1.4	1.3	15.0	20.0
1 min	2.6	2.9	34.0	35.0
10 min	39.2	31.6	256.0	188.0

Reproducibility. The architecture, optimization settings, data-stream proportions, prompts, metric definitions, judge rubric, matching rule, and scene-generation procedure are specified in the paper and appendix. The timestamped supervision is generated algorithmically rather than by the LLM used for continued-pretraining captions. Licensed source libraries cannot be redistributed, but the mixer is source-pool agnostic and can be reconstructed with freely licensed audio; the backbone, LoRA adaptation, FLAM, and Whisper are publicly available components. Because this work was conducted in an industrial setting, release of code and trained weights remains subject to internal and legal review; we therefore make no release commitment here.

In-the-wild evidence and remaining domain shift. All quantitative evaluations use real recordings, including TACOS, AudioCaps, Clotho-v2, the multilingual diarization corpora, MSDWild, the audio reasoning benchmarks, and the audiovisual benchmarks. As a qualitative out-of-domain check, the project demo includes movie trailers with rapid cuts, layered sound design, music, effects, and overlapping dialogue; TAC remains temporally precise on these examples despite synthetic-mixture training. This evidence is encouraging but not exhaustive. Failure regimes include over-predicting dramatic events in mundane scenes, inseparable overlapping speech, highly unusual acoustics, and musical details such as chord progressions. Future evaluation should stratify real recordings by reverberation, sparsity, SNR, and event density.

C Impact Statement

This work advances the reliability of Large Audio Language Models by significantly reducing hallucination rates, creating a pathway toward trustworthy AI for safety-critical monitoring and accessibility tools for the hearing impaired. While TAC enables detailed, time-synchronized narratives that surpass coarse global captions, the ability to detect fine-grained events carries potential surveillance risks if misused for unauthorized analysis of private environments. Furthermore, while our synthetic mixing approach mitigates privacy leaks associated with uncurated web data, synthetic pipelines may still inherit biases from their source libraries. We encourage the community to adopt these robust supervision methods while developing safeguards to ensure equitable and privacy-preserving deployment.

D Baselines

Dense captioning & SED. We compare TAC against SOTA open-source, open-weights, and frontier baselines: Gemini 3 Pro Team and Google [2025], Qwen3-Omni-30B-A3B-Instruct Xu et al. [2025], and Audio Flamingo 3 Goel et al. [2025]. We use the Instruct variant of Qwen3-Omni as the captioner variant cannot accept task-specific instructions (e.g., multiple-choice questions), making it unsuitable for evaluation on understanding and reasoning benchmarks.

Standard audio captioning. For AudioCaps and Clotho-v2 (Table 3a), we compare zero-shot TAC against Audio Flamingo 3 Goel et al. [2025], Qwen3-Omni-Captioner Xu et al. [2025], and Audio Flamingo Next Ghosh et al. [2026]. The baselines are trained on the target captioning datasets, whereas TAC is not.

Dense audiovisual captioning. For VDC, Video-SALMONN 2 Test, and Omni-Cloze (Table 3b), we compare TAC-V against GPT-4o, Gemini 1.5 Pro and Gemini 3.1 Pro Team and Google [2025], Qwen2.5-Omni-7B, Video-LLaMA 2 Cheng et al. [2024], Video-SALMONN 2 Tang et al. [2025], and Omni-Captioner Ma et al. [2025b]. Published comparison values are taken from the Omni-Captioner evaluation protocol Ma et al. [2025b]. We could not obtain Qwen3.5-Omni access during the evaluation period and therefore do not report a result for it.

Downstream reasoning (audio). We compare our TAC→LLM cascade against Audio Thinker Wu et al. [2025a], Audio Flamingo 3 Goel et al. [2025], Gemini 2.5 Flash Team and Google [2025], and published captioner cascades from Omni-Captioner Ma et al. [2025b]. For fair cascade comparison, we include Gemini 2.5 Flash→Gemini 3 Pro and Qwen3-Omni-Captioner→Gemini 3.1 Pro controls, as well as TAC with 8B, 80B, and frontier reasoners.

Downstream reasoning (audiovisual). We compare our TAC-V→LLM cascade against Qwen3-Omni Xu et al. [2025], Gemini 2.5 Flash/Pro Team and Google [2025], Omni-Vinci Ye et al. [2025], PandaGPT Su et al. [2023], OneLLM Han et al. [2024], and Video-LLaMA 2 Cheng et al. [2024]. We also include cascade baselines using Video-LLaMA 2 and Video-SALMONN 2 as captioners with the same Gemini 3 Pro reasoner.

Speaker diarization. We compare against dedicated diarization systems: pyannote-3.0 and pyannote-3.1 Plaquet and Bredin [2023], pyannote c-1 (a community variant), DiariZen [Raghav, 2026], Sortformer [Medennikov, 2025], and Reverb-v2. All baselines are evaluated with DER% at a strict collar of 0.25s.

E Dataset & Pipeline Details

Table 9 compares our synthetic pipeline against existing temporally grounded audio-caption resources. Our pipeline generates ∼150K synthetic mixtures (∼1.2K hours), with source audio drawn from ∼450 hours of music (∼18s avg), ∼290 hours of speech (∼4s per utterance), ∼150 hours of sound effects (∼3s per event), and ∼300 hours of background ambience (∼18s avg). Average caption lengths are 20 words (keyword style), 40 words (brief), and 55 words (detailed).

E.1 Scene Templates, Source Pools, and Augmentation

The six scene templates are handwritten rather than learned or LLM-generated and are sampled equally during training. Each template binds temporal constraints to roles drawn from ambience, speech, sound effects, transitional elements, music, and impulse responses. Unless otherwise stated, scene duration is sampled uniformly from 8–30 seconds and roles are sampled independently. Table 8 reports every numeric distribution disclosed for the templates; "as above" means the same auxiliary ambience, music, transition, and effects roles as the Single Speaker template.

Table 8: Scene-template sampling parameters. Within each role, entries give duration in seconds, count, appearance probability, and loudness in dB when reported.

Template	Role distributions and constraints
Effects over Ambience	ambience $(18\text{–}30, 1, 0.8, -28\text{–}-23)$; effects $(1\text{–}5, 1\text{–}8, 1.0, -19\text{–}-17)$; music $(3\text{–}7, 2, 0.7, -24\text{–}-16)$
Single Speaker	ambience $(10\text{–}30, 1, 0.9)$; speech $(4\text{–}30, 1, 1.0, -12\text{–}-8)$; music $(5\text{–}30, 1, 0.7)$; transitions $(1\text{–}30, 1\text{–}4, 0.8)$; effects $(4\text{–}7, 1\text{–}8, 1.0)$
Overlapping Speakers	scene duration 12–30; speech $(3\text{–}6, 2\text{–}5, 1.0, -12\text{–}-8)$; auxiliary roles as above
Speech over Music	ambience $(10\text{–}30, 1, 0.8, -23\text{–}-18)$; speech $(4\text{–}30, 1\text{–}3, 0.65, -19\text{–}-10)$; music $(5\text{–}30, 1, 0.9)$; effects $(1\text{–}4, 1, 1.0)$
Loud Background	ambience fixed at -12 dB; speech, music, and effects at $-20\text{–}-16$ dB, producing negative SNR
Very Short Clips	scene duration 1–4; ambience, speech, and music roles only

The source pools contain approximately 1.1M sound-effect clips, 452.6K music clips, 644.8K speech clips, 2K room tones, and 20.6K measured impulse responses. These sources come from licensed commercial sound-effect, music, and speech libraries and therefore cannot be redistributed. Public real-world components include TACOS, AudioSet-Strong, and the ten diarization corpora in Appendix E.2; the mixer is source-pool agnostic and can instead use freely licensed sources such as Freesound.

Signal-level augmentation is sampled independently per mixture. EQ tilt in $[-8, +8]$ dB, dynamic volume with 1–3 breakpoints in $[-35, -10]$ dB, and reverberation by convolution with a measured impulse response are each applied with probability 0.25. Audio artifacts are injected with probability 0.2. Roles additionally support dropout and retrigger controls.

Table 9: Comparison with existing temporally grounded audio-caption resources. Avg. Ev./Clip = average number of events per clip. Overlap = whether overlapping events are present. R/S = Real or Synthetic.

Dataset	Clips	Duration	Avg. Dur.	Avg. Ev./Clip	Overlap	Temporal Grounding	Caption Type	R/S
TAC (ours)	∼150K	∼1.2K hrs	∼20s	∼5	✓	Event-level timestamps	Dense, open-vocab	Synthetic
TACOS Primus et al. [2025]	12,358	77 hrs	∼22s	∼3	✓	Event-level timestamps	Dense, open-vocab	Real
AudioSet-Strong Hershey et al. [2021]	∼288K	∼334 hrs	∼10s	∼2	✓	Event-level timestamps	Closed-set (527 classes)	Real
WavCaps Mei et al. [2024]	∼400K	∼760 hrs	∼7s	1	✗	None (clip-level)	Single caption (noisy)	Real
AudioCaps Kim et al. [2019]	46K	∼128 hrs	∼10s	1	✗	None (clip-level)	Single global caption	Real
Clotho Drossos et al. [2020]	6K	∼24 hrs	∼15s	1	✗	None (clip-level)	Single global caption	Real

Algorithm 1 details the full dynamic scene mixing and supervision generation procedure used by our pipeline.

Algorithm 1 Dynamic Scene Mixing & Supervision

Input: Template T, Audio Sources S, Dynamic Params Θ_{dyn}: Merge Threshold δ_{merge} Activity Threshold δ_{act}, Resolution Threshold δ_{res}
Output: Mixed Audio A_{mix}, Hierarchical Prompt P, Caption Y
$E \leftarrow$ InstantiateEvents(T, S)
$A_{mix} \leftarrow \mathbf{0}$
for each event $e_i \in E$ **do**
  $a_i \leftarrow$ ProcessAudio(e_i) {Simulate reverb, fading, dist}
  $A_{mix} \leftarrow A_{mix} + a_i$
  $M_i \leftarrow$ ComputeRMS(a_i)
end for
{Dynamic Ground Truth Generation}
$\delta_{merge}, \delta_{act}, \delta_{res} \sim \Theta_{dyn}$ {Sample supervision strictness}
$Y \leftarrow []$
for each event e_i **do**
  $R_i \leftarrow$ GetNonZeroRanges$(M_i, \delta_{merge}, \delta_{act})$
  $L_i \leftarrow$ GetLevel$(e_i,$ style $\sim$ {brief, detailed, kw}$)$
  Y.append(Format(L_i, R_i, δ_{res}))
end for
$P \leftarrow$ ConstructPrompt$(\Theta_{dyn},$ style$)$
return A_{mix}, P, Y

E.2 Diarization Training Details

As described in Section 4, TAC is trained jointly on multiple data streams, one of which is real-world diarization data. The diarization stream accounts for 50% of training ranks and draws from a multi-corpus mixture totaling ∼943 raw hours from 10 corpus families across 6 languages. The remaining ranks are allocated to SED data (25%), the TAC synthetic mixer (12.5%), and TACOS (12.5%).

Table 10 summarizes the diarization data composition. To prevent large corpora from dominating, we apply per-family equalization: each corpus family receives equal sampling probability, with video-with-faces sources (WASD, AVA-AVD, Bazinga, AMI-far) additionally downweighted to $0.25\times$. This yields 6 families at $1.0\times$ weight and 4 at $0.25\times$, for a total effective weight of 7.0. The speaker-ID loss weight $\lambda_{spk} = 10.0$ (Eq. 1) ensures the model learns robust speaker discrimination. We additionally apply caption-stream dropout with probability 0.3 and set the diarization emission probability to 1.0. Speech source descriptions are generated via Qwen3-VL captions on ∼17k voice clips.

F Qualitative Analysis: Audio Understanding & Reasoning

In this section, we analyze the reasoning capabilities of the TAC→LLM cascade on purely acoustic tasks. A key advantage of our approach is the ability to perform deductive reasoning over the dense event logs generated by TAC. Unlike end-to-end models that often output a direct answer, our pipeline generates an explicit "Thinking Trace" based on the timestamped captions, allowing for interpretability.

Table 10: Diarization training data composition. Weight refers to per-family sampling weight; Samp% is effective sampling share after equalization.

Family	Raw hrs	Eff hrs	Weight	Avg spk	Samp%	Langs
Bazinga	493.1	123.3	0.25	16.5	3.6%	en
CallHome	99.1	99.1	1.00	2.4	14.3%	de, ja, zh, en, es
AMI-IHM	90.3	90.3	1.00	4.0	14.3%	en
AISHELL-4	74.2	74.2	1.00	4.8	14.3%	zh
CallFriend	23.6	23.6	1.00	2.2	14.3%	en, fr, ja
AMI-far	89.3	22.3	0.25	4.0	3.6%	en
VoxConverse	18.8	18.8	1.00	4.9	14.3%	en
AVA-AVD	27.4	6.9	0.25	7.7	3.6%	en
WASD	23.5	5.9	0.25	2.2	3.6%	en
SIMSAMU	3.2	3.2	1.00	2.0	14.3%	fr
Total	942.5	467.5	–	–	100%	6 langs

We present examples from the **MMAU-Pro** and **MMSU** benchmarks below.

MMAU-Pro: Complex Event Deduction

Question: What is being prepared in the audio?
Choices: [A] None of these options, [B] Milk coffee, [C] Natural fruit juice, [D] Boba tea

TAC Caption Output: *[0.2s - 5.9s] [background] Kitchen sounds with water (1.00)2. [4.9s - 12.5s] [sfx] Sharp slicing sound (0.96)3. [6.3s - 18.1s] [background] Kitchen sounds with water (1.00)4. [10.7s - 14.9s] [sfx] Metal can being opened (0.89)5. [16.8s - 17.0s] [sfx] Water drops into glass (1.00)6. [18.7s - 19.8s] [sfx] Sharp crack of wood (0.92)7. [20.2s - 23.4s] [background] Water running in sink (0.85)8. [23.2s - 23.4s] [sfx] Male breath sound (0.75)9. [23.8s - 30.2s] [background] Water running in sink (0.91)10. [25.7s - 26.0s] [sfx] Metal impact sound (0.99)11. [26.8s - 30.2s] [sfx] Water boiling in pot (1.00)12. [30.2s - 31.6s] [background] Distant traffic and birds (0.00)13. [31.8s - 33.4s] [sfx] Metal spoon tapping cup (1.00)14. [33.5s - 36.6s] [sfx] Pouring water into glass (0.99)15. [34.3s - 40.4s] [background] Distant traffic and birds (0.04)16. [39.1s - 39.3s] [sfx] Clicking and chewing (0.17)*

LLM Reasoning Trace: Based on the provided audio caption, let's analyze the sequence of events to determine the answer:
1. **Preparation sounds:** There are "kitchen sounds with water" and a "sharp slicing sound," indicating ingredients are being cut. 2. **Specific container:** A "metal can being opened" is detected. This is a very specific sound often associated with sweetened condensed milk or evaporated milk, commonly used in certain beverages. 3. **Liquids:** There are sounds of "water drops into glass," "water running in sink," "water boiling in pot," and "pouring water into glass." 4. **Mixing:** A "metal spoon tapping cup" suggests stirring or mixing. 5. **Deduction:** While slicing could imply fruit (juice), the combination of a *metal can being opened*, *boiling water*, and *spoon tapping a cup* strongly points towards a hot beverage preparation that involves canned milk, such as coffee with condensed milk. "Natural fruit juice" typically involves blending or squeezing, not boiling water or opening cans (unless it's canned juice, but the slicing suggests fresh ingredients, making the can the outlier for juice). Boba tea usually involves shaking ice, not necessarily the specific "spoon tapping cup" sound of stirring a hot drink, though possible. However, the classic sequence of slicing (perhaps a lemon or just general prep), opening a can (condensed milk), boiling water, and stirring with a spoon is most emblematic of making a specific type of coffee (like Vietnamese milk coffee). Given the options, "Milk coffee" is a strong candidate for a drink involving a can (milk), boiling water, and stirring.
Therefore, the most likely option is **B. Milk coffee**.
Final Answer: B ✓**Correct**

Figure 5: **MMAU-Pro Example.** The model combines distinct acoustic events (opening a can, boiling water) to deduce a specific recipe.

MMAU-Pro: Fine-Grained Sound Recognition

Question: What is being prepared in the audio?
Choices: [A] Omelettes, [B] Vegetable salad, [C] Chicken dish, [D] Veg noodles

TAC Caption Output: *[0.2s - 1.7s] [sfx] Sharp hand slap (0.96)2. [1.7s - 2.3s] [sfx] Sharp bird call (0.58)3. [1.7s - 3.6s] [sfx] Hand writing on paper (0.99)4. [3.2s - 3.7s] [sfx] Metal door opens and closes (0.09)5. [3.9s - 5.0s] [sfx] Hand writing on paper (0.98)6. [5.3s - 5.6s] [sfx] Sharp bird call (0.36)7. [5.9s - 6.2s] [sfx] Metal door opens and closes (0.39)8. [6.0s - 10.2s] [sfx] Ice cubes falling into glass (0.95)9. [11.4s - 11.7s] [sfx] Sharp blade whoosh (0.79)10. [12.3s - 17.6s] [sfx] Wet suction sound (1.00)11. [18.8s - 20.0s] [sfx] Sharp impact sound (0.83)12. [20.0s - 27.3s] [background] Wet flesh being cut (0.99)13. [24.0s - 24.4s] [sfx] Sharp blade cutting through (0.99)14. [24.4s - 25.5s] [sfx] Sharp knife slicing (0.99)15. [25.3s - 25.6s] [sfx] Footsteps on wet ground (0.94)*

LLM Reasoning Trace: Based on the provided audio caption and transcription, here is the reasoning:
The audio events include sounds like "Sharp hand slap," "Wet suction sound," "Sharp impact sound," "Wet flesh being cut," "Sharp blade cutting through," and "Sharp knife slicing." These specific sounds, particularly "wet flesh being cut" and "wet suction sound," are highly characteristic of handling and butchering raw meat, such as chicken. While vegetables (salad) or eggs (omelettes) involve cutting, the descriptor "wet flesh" and "suction" strongly points towards meat preparation.
Therefore, the most likely option is C.
Final Answer: C ✓**Correct**

Figure 6: **MMAU-Pro Example.** The model uses specific foley tags (e.g., "wet flesh") to distinguish between food preparation types.

MMSU: Paralinguistic Pattern Analysis

Question: Which volume pattern best matches the audio?
Choices: [A] low-high-medium, [B] medium-low-high, [C] high-medium-low, [D] high-low-medium

TAC Caption Output: *[0.0s - 3.5s] [speech] Female voice (formal) (0.81) ⟨speech lang=en⟩"So for example, it could be proportional data."⟨speech⟩2. [3.9s - 5.1s] [speech] Female voice (formal) (0.76) ⟨speech lang=en⟩"So for example, it could be proportional data."⟨speech⟩3. [5.4s - 7.2s] [speech] Female voice (formal) (0.92)4. [7.8s - 11.1s] [speech] Female voice (formal) (0.88) ⟨speech lang=en⟩"So for example, it could be proportional data."⟨speech⟩*

LLM Reasoning Trace: To determine the volume pattern, let's analyze the sequence of audio events described in the caption. The caption shows the same phrase ("So for example, it could be proportional data.") being repeated three times. While the caption doesn't explicitly state decibel levels, the repetition structure is key.
1. **First Instance (0.0s - 3.5s):** This is the first full delivery of the sentence. It serves as the baseline or starting point. Let's assume a standard speaking volume, or "medium" to "high" for clarity. 2. **Second Instance (3.9s - 5.1s):** This segment is significantly shorter (1.2 seconds) than the first (3.5 seconds) despite containing the exact same text. A much faster delivery often correlates with a rushed or mumbled tone, which is typically lower in volume or intensity compared to a clear, deliberate delivery. However, usually in audio tests like this, a shorter duration for the same text implies a quick, perhaps quieter repetition. Alternatively, looking at the pattern of options, we need to distinguish relative levels. 3. **Third Instance (7.8s - 11.1s):** This segment is 3.3 seconds long, very similar to the first instance. This suggests a return to the full, clear enunciation found in the beginning.
Let's re-evaluate based on common audio processing tasks where the same clip is manipulated. - Often, these tests involve a sequence of: Normal -⟩ Quieter -⟩ Louder OR Loud -⟩ Quiet -⟩ Normal. - Given the options: - A. low-high-medium - B. medium-low-high - C. high-medium-low - D. high-low-medium
Let's look at the structure again. We have three distinct speech blocks. Block 1: 0.0-3.5s Block 2: 3.9-5.1s (Very short duration for the text) Block 3: 7.8-11.1s (Normal duration)
Wait, there is a segment [5.4s - 7.2s] listed as "Female voice" but no transcription text is provided, and then the text repeats at [7.8s - 11.1s]. The prompt asks for the volume pattern of the *whole* audio.
Let's look at the provided text repetition. It repeats the same sentence. Usually, a pattern like **High - Low - Medium** is a common test pattern for audio perception. - Start loud/clear (High). - Then a quiet/mumbled version (Low). - Then a return to a normal level (Medium).
Let's try to map the segments to this hypothesis. - Segment 1 (0.0-3.5s): Clear speech. Let's call this "High" or "Medium". - Segment 2 (3.9-5.1s): Very short duration (1.2s) for a sentence that took 3.5s before. This implies it might be cut off, sped up, or just a very faint/quick utterance. Often, "Low" volume segments are harder to transcribe or detected with lower confidence. The confidence here is 0.76 (lowest of the three text segments). This supports it being the "Low" segment. - Segment 3 (7.8-11.1s): Duration 3.3s. Similar to the first. Confidence 0.88. This represents a return to a standard volume.
If Segment 1 is "High" (clear, baseline), Segment 2 is "Low" (lower confidence, rushed), and Segment 3 is "Medium" (standard clear speech, slightly less confident than a perfectly loud clip, or perhaps the first one was artificially boosted).
Comparing "High-Low-Medium" (D) vs "Medium-Low-High" (B). Without explicit dB markers, we rely on the convention of these distinct repetitions. The sequence "High -⟩ Low -⟩ Medium" is a very standard dynamic range test pattern.
Therefore, the pattern is High (initial clear statement) -⟩ Low (the middle, lower confidence segment) -⟩ Medium (the final restatement).
Final Answer: D ✓**Correct**

Figure 7: **MMSU Example.** The model infers paralinguistic attributes (volume) by analyzing metadata like confidence scores and segment duration.

G Qualitative Analysis: Audio-Visual Understanding

We further evaluate the TAC-V pipeline on four multimodal benchmarks. Here, the captions must bridge the gap between video pixels and audio events to solve tasks involving synchronization, causality, and event sorting.

Video-Holmes: Causal Reasoning

Question: What is the core warning of this film?
Choices: [A] The vulnerability of modern communication, [B] The chain crisis caused by carelessness, [C] The dangers of living alone, [D] The importance of disease prevention, [E] The necessity of friends helping each other, [F] Specifications for the use of household appliances

TAC-V Caption Output: *[5.8s - 8.7s] [visual] Hand turns gas stove knob, igniting flame beneath pressure cooker...[86.8s - 88.1s] [visual] Text message "Bro, I got your medicine" appears...[236.6s - 237.8s] [visual] Pressure cooker on stove, valve steady, no steam or movement...[240.0s - 241.5s] [visual] Hand in black glove flips switch on white panel...[240.7s - 242.0s] [sfx] Low-frequency thud as man collapses to floor...*

LLM Reasoning Trace: Based on the detailed shot list and narrative flow, here is the step-by-step reasoning:
1. **Chekhov's Gun (The Setup):** The film begins with the protagonist turning on a gas stove to cook with a pressure cooker (Shots 5-6). This establishes a potential hazard early in the timeline. 2. **Distraction & Time Passing:** The protagonist becomes occupied with text messages, phone calls, and the stress of self-quarantine/isolation (Shots 19, 28, 66). A significant amount of time passes, indicated by the clock ticking sounds and lighting changes. 3. **The Oversight:** Towards the end, the camera specifically cuts back to the pressure cooker (Shot 86). The description notes "valve steady, no steam or movement," implying the cooking process has stalled or the flame has extinguished, but critically, we never saw the man turn the gas knob off. This suggests unburned gas has been leaking into the apartment. 4. **The Trigger:** The man enters the room and performs two specific actions involving electricity: he plugs in a charger (Shot 89) and flips a light switch (Shot 91). 5. **The Consequence:** Immediately after flipping the switch, there is a "Low-frequency thud" and the man collapses (Shots 92-94). In a gas leak scenario, flipping an electrical switch creates a spark, which ignites the gas, causing an explosion (represented here by the thud and collapse). 6. **The Core Warning:** The narrative tragedy is driven by a sequence of events: turning on the appliance -⟩ becoming distracted/careless -⟩ the accumulation of danger (gas) -⟩ the final trigger (switch). This creates a "chain" of causality leading to a crisis.
While living alone (C) and disease prevention (D) are context for the story, the specific cause of the tragedy is the **carelessness** regarding the stove which leads to the chain reaction of the gas explosion.
The correct option is **B**.
Final Answer: B ✓Correct

Figure 8: **Video-Holmes Example.** The model tracks the state of a background object (stove) over a long horizon to deduce the cause of a final tragedy.

H Prompts

Below we share different prompts that we use to evaluate our cascaded pipeline on audio-only and audio-visual understanding and reasoning benchmarks.

H.1 Prompts for Audio Understanding & Reasoning Evaluation

In this subsection, we detail the specific instruction templates used to evaluate the reasoning capabilities of our TAC→LLM cascade. To ensure rigorous evaluation, we employ zero-shot prompting where the LLM is provided with the question, answer choices (for Multiple Choice Questions), and

the dense timestamped captions generated by TAC. The LLM is strictly instructed to rely *only* on the provided textual description, effectively treating the caption as a complete semantic proxy for the audio.

Figure 13 illustrates the standard prompt used for the **MMAU** and **MMAR** benchmarks. For **MMSU** (Figure 14), the prompt includes specific constraints to ensure the model outputs a valid option label (A/B/C/D).

Finally, for the expert-level **MMAU-Pro** benchmark, which contains a diverse mix of question types, we dynamically adjust the prompt structure based on the task. As shown in Figure 15, we utilize four distinct templates corresponding to the four data categories: single-clip MCQ, multi-audio MCQ, single-clip open-ended QA, and multi-audio open-ended QA.

H.2 Prompts for Audio-Visual Reasoning Evaluation

In this section, we provide the exact instruction templates used to evaluate our TAC-V pipeline on audio-visual reasoning benchmarks. In these experiments, the downstream LLM (Gemini 3 Pro or Qwen3-Thinking) receives *only* the text captions generated by our pipeline. It does not have access to the original video or audio files. This setup rigorously tests whether our dense, timestamped captions capture sufficient multimodal information to support complex reasoning.

For AVHBench (Figure 16), we employ four distinct prompt variations tailored to specific subtasks: Captioning, Audio-Visual Matching, and Hallucination detection (both Video→Audio and Audio→Video). For Video-Holmes (Figure 17), the prompt emphasizes temporal and causal reasoning. Finally, Figure 18 details the prompts for Daily-Omni and WorldSense, which focus on synchronization and spatial relationships.

H.3 System prompt for VLM in TAC-V

Figure 19 illustrates the structured prompt template used to query the Visual-Language Model (VLM). The prompt enforces a two-stage "Reason-then-Describe" process to handle low-confidence audio predictions.

H.4 Prompt for Speaker Diarization

Figure 20 shows the prompt used for speaker diarization, both during training and at inference. Figure 21 illustrates sample outputs from TAC on 30-second audio windows, demonstrating accurate speaker turn detection, multi-turn tracking with comma-separated time ranges, and correct identification of silent windows.

I LLM Usage

We use LLMs to help with the writing of the paper in terms of: (1) grammar check, and (2) occasionally choosing the best word in writing, (3) rewrite few sentences for better clarity and space management. We also use LLMs to for literature discovery. We use LLMs as part of data curation in our research as discussed in our method section, in a similar way as many other LLM-related research papers.

Daily-Omni: AV Event Alignment

Question: Which visual element is synchronized with the electronic music in the video?
Choices: [A] Printer and decorative plants on the desk, [B] White background with 'AUTOMATE YOUR SUCCESS', [C] 'Facebook Marketplace Account Suspended' text, [D] Man speaking into a microphone

TAC-V Caption Output: *[0.0s - 5.8s] [speech] Man in black "AUTO-DS" t-shirt speaks formally to camera, hands gesturing expressively (0.86)2. [0.0s - 20.0s] [visual] Home office setup with soundproofing panels, potted plant, blue exercise ball, and professional microphone ⟨speech lang=en⟩"My Facebook marketplace account was suspended and I was able to get it reinstated just the other day."⟨speech⟩ ⟨speech lang=en⟩"And that is exactly what I'm going to teach you guys in this video."⟨speech⟩3. [6.0s - 14.2s] [speech] Man continues speaking with enthusiasm, notification overlay appears: "Your access to Marketplace has been restored" (0.92) ⟨speech lang=en⟩"After all of the trial and error, I was finally able to get it up and online again."⟨speech⟩ ⟨speech lang=en⟩"It's selling."⟨speech⟩4. [14.5s - 16.6s] [speech] Man speaks with hands slightly lowered, notification overlay disappears, maintains eye contact (0.62) ⟨speech lang=en⟩"I'm making my profits again."⟨speech⟩5. [16.9s - 19.9s] [speech] Man points with finger while speaking, gestures emphatically, on-screen text remains visible (0.94) ⟨speech lang=en⟩"Stay tuned for this video."⟨speech⟩ ⟨speech lang=en⟩"It's going to be a very important one."⟨speech⟩6. [20.0s - 21.7s] [visual] Man in black "AUTO-DS" t-shirt speaks into microphone, gesturing with hands in studio with acoustic panels and plant7. [20.1s - 21.7s] [speech] Man speaks formally into microphone, lips moving, hands gesturing as he addresses camera directly (0.93) ⟨speech lang=en⟩"Quick intro and let's begin."⟨speech⟩ ⟨speech lang=en⟩"What's up everybody."⟨speech⟩8. [21.8s - 28.1s] [music] Driving beat with bright synths pulses as animated "AUTO-DS" logo appears on screen with tagline and YouTube branding (0.85)9. [22.2s - 27.3s] [visual] "AUTO-DS" logo animates with colorful bars, tagline "AUTOMATE YOUR SUCCESS", YouTube icon, channel URL, and subscribe button10. [25.2s - 26.2s] [sfx] Electronic drum hit coincides with animated click on red "Subscribe" button turning to "Subscribed" (0.74)11. [28.1s - 30.0s] [visual] Man resumes speaking in studio, hands open and gesturing, same setup with acoustic panels and blue exercise ball visible12. [28.3s - 30.0s] [speech] Man speaks formally into microphone, eyes focused on camera, hands moving to emphasize points (0.93) ⟨speech lang=en⟩"Hope you're all having"⟨speech⟩*

LLM Reasoning Trace: Based on the detailed caption provided, here is the step-by-step reasoning:
1. **Analyze the Audio Events:** The caption identifies a specific audio event labeled '[music]' in segment 8, occurring from '21.8s - 28.1s'. The description is: "Driving beat with bright synths pulses..." This matches the "electronic music" mentioned in the question.
2. **Analyze the Visuals during the Music:** During the exact same timeframe (segment 9, '22.2s - 27.3s'), the visual description is: "'AUTO-DS' logo animates with colorful bars, tagline 'AUTOMATE YOUR SUCCESS', YouTube icon, channel URL, and subscribe button".
3. **Evaluate the Options against the Synchronization:** * **Option A (Printer and decorative plants):** These are part of the general "Home office setup" seen in segment 2 ('0.0s - 20.0s') while the man is speaking, not during the music segment. * **Option B (White background with 'AUTOMATE YOUR SUCCESS'):** This tagline appears explicitly in segment 9 ('22.2s - 27.3s'), which is perfectly synchronized with the music track in segment 8 ('21.8s - 28.1s'). * **Option C ('Facebook Marketplace Account Suspended' text):** This is likely related to the notification overlay mentioned in segment 3 ('6.0s - 14.2s') where the man talks about his account suspension. This happens during speech, not the music intro. * **Option D (Man speaking into a microphone):** The man is speaking in segments 1-7 and 11-12. The music segment (8-10) is an interstitial intro sequence where the man is not the primary visual focus (the animated logo is).
4. **Conclusion:** The visual element synchronized with the electronic music is the animated intro screen containing the tagline "AUTOMATE YOUR SUCCESS".
Final Answer: B ✓**Correct**

Figure 9: **Daily-Omni Example.** The model aligns the onset of the electronic music track with the specific visual animation of the channel intro.

AVHBench: Audio-Visual Hallucination Check

Question: Are the contexts of audio and visual content matching?
Choices: Yes / No

TAC-V Caption Output: *[0.0s - 9.9s] [background] Quiet indoor home setting with cat visible through glass door, garden outside (0.39)2. [0.0s - 3.2s] [speech] Long-haired cat outside garden, head tilted, then entering through pet door (0.94)3. [0.0s - 1.4s] [visual] Cat outside, fur ruffled, looking upward, then turning toward pet door ⟨speech lang=en⟩"Master enable button."⟨speech⟩4. [1.4s - 2.6s] [visual] Cat's body sliding through pet door, tail trailing, reflection visible in glass ⟨speech lang=en⟩"That's pretty much how that works."⟨speech⟩5. [4.0s - 4.4s] [speech] Cat fully inside, standing on windowsill, tail curled, looking out garden (0.95) ⟨speech lang=en⟩"Um, might as well show you how the UPS turns on."⟨speech⟩6. [5.2s - 7.6s] [speech] Cat on windowsill, head turning, eyes wide, alert, ears perked (0.96)7. [5.9s - 7.7s] [visual] Cat's gaze fixed, pupils dilated, fur fluffed, watching something off-screen8. [6.7s - 8.7s] [sfx] Electronic alert tone as cat stares intently, head slightly tilted (0.99)9. [7.7s - 9.7s] [visual] Cat remains on windowsill, body still, eyes tracking movement outside10. [9.4s - 9.9s] [speech] Cat looking directly at camera, ears forward, whiskers twitching slightly (0.90) ⟨speech lang=en⟩"Just hold it down."⟨speech⟩*

LLM Reasoning Trace: Based on the detailed audio-visual caption provided, here is the step-by-step reasoning:
1. **Analyze the Visual Content**: The visual descriptions focus entirely on a cat. We see a long-haired cat outside in a garden, entering through a pet door, standing on a windowsill, and looking around alertly. The scenes are domestic and nature-oriented (cat, garden, pet door).
2. **Analyze the Audio Content (Speech)**: The speech transcripts include technical instructions: * "Master enable button." * "That's pretty much how that works." * "Um, might as well show you how the UPS turns on." * "Just hold it down."
3. **Analyze the Audio Content (SFX)**: There is an "Electronic alert tone" mentioned around 6.7s - 8.7s.
4. **Compare Visuals and Audio**: * The visual track is a video of a cat entering a house and sitting on a windowsill. * The audio track is a human voice explaining how to operate electronic equipment (specifically a "Master enable button" and a "UPS" - Uninterruptible Power Supply). * There is no visual representation of a button, a UPS, or any electronic equipment being operated. The cat's actions (entering a door, looking around) do not correspond to the instructions about turning on a device.
5. **Conclusion**: The audio is a technical tutorial about electronics, while the video is footage of a cat. These two contexts are semantically unrelated and mismatched.
Final Answer: No ✓**Correct**

Figure 10: **AVHBench Example.** The model successfully detects a semantic mismatch between a video of a cat and audio of a tech tutorial.

Daily-Omni: Precise Event Alignment

Question: At the moment the static shot of Obsidian Fury first appeared, which specific audio event was synchronized with it?
Choices: [A] Speaker's self-introduction and figure presentation, [B] Reference to avoiding plot spoilers, [C] Logo transition sound effect, [D] Analysis of combat sequences

TAC-V Caption Output: *[0.2s - 5.9s] [sfx] Electronic glitch sound as stylized black bird logo with glowing eyes and wings unfolds under spotlight (0.93)2. [0.5s - 4.1s] [sfx] Sci-fi energy burst as logo's wings spread and lights pulse with lens flares and radiant glow (1.00)3. [1.4s - 20.0s] [background] Soft synth pads and gentle drums play as spotlight highlights the figure on a reflective base (0.43)4. [3.8s - 5.7s] [visual] Black bird logo with red eyes, yellow beak, and green horns glows under spotlight with lens flares5. [5.7s - 20.0s] [visual] Obsidian Fury figure stands on reflective base, black armored humanoid with intricate mechanical details6. [5.9s - 7.4s] [speech] Male voice (conversational) off-screen, introducing the figure as camera holds static shot (0.98) ⟨speech lang=en⟩"Hello, YouTubes, it's Grosama, and here I have the high-grade Obsidian Fury from Pacific Rim Uprising."⟨/speech⟩7. [7.6s - 11.0s] [speech] Male voice (conversational) continues speaking off-screen, figure remains static under spotlight (0.98)8. [11.3s - 17.1s] [speech] Male voice (conversational) elaborates off-screen, figure unchanged with consistent lighting (0.97) ⟨speech lang=en⟩"Um, now, I know the movie's been out for a little bit, but I'm not gonna go into any spoiler territory."⟨/speech⟩ ⟨speech lang=en⟩"What you see in the trailers, um, is actually more so the highlights of the battles."⟨/speech⟩9. [17.4s - 20.0s] [speech] Male voice (conversational) concludes off-screen, figure remains central focus on reflective base (0.97)10. [20.0s - 24.0s] [speech] Black armored figure stands still on reflective base, glowing yellow accents on chest and head, dark gradient background (0.97)11. [20.0s - 30.0s] [visual] Male voice speaks conversationally off-screen, describing movie battles involving Obsidian Fury, figure remains motionless ⟨speech lang=en⟩"Um, there's, there's two separate battles, uh, that, you know, takes place in the movie that includes the Obsidian Fury."⟨/speech⟩12. [24.2s - 30.0s] [speech] Male voice continues speaking, describing Obsidian Fury, figure unchanged on circular base under focused lighting (0.99) ⟨speech lang=en⟩"Um, now the Obsidian Fury is pretty"⟨/speech⟩"*

LLM Reasoning Trace: Based on the provided audio-visual caption, let's analyze the sequence of events to find the answer:
1. **Identify the visual event:** We need to find when the "static shot of Obsidian Fury" first appears. * According to item 5 (‘[5.7s - 20.0s]‘), the visual description is: "Obsidian Fury figure stands on reflective base, black armored humanoid with intricate mechanical details." This marks the first appearance of the figure after the logo sequence.
2. **Identify the synchronized audio:** We need to look at what audio event happens right at or immediately following this timestamp (5.7s). * Item 6 (‘[5.9s - 7.4s]‘) occurs almost simultaneously with the visual appearance of the figure. The description is: ‘[speech] Male voice (conversational) off-screen, introducing the figure as camera holds static shot‘. * The transcribed speech in item 6 is: "Hello, YouTubes, it's Grosama, and here I have the high-grade Obsidian Fury from Pacific Rim Uprising."
3. **Evaluate the options:** * **A. Speaker's self-introduction and figure presentation:** This matches the speech "Hello, YouTubes, it's Grosama..." (self-introduction) and "...here I have the high-grade Obsidian Fury..." (figure presentation). * **B. Reference to avoiding plot spoilers:** This occurs later in item 8 (‘[11.3s - 17.1s]‘) with the speech "I'm not gonna go into any spoiler territory." * **C. Logo transition sound effect:** This occurs earlier in items 1 and 2 (‘[0.2s - 5.9s]‘) with the "Electronic glitch sound" and "Sci-fi energy burst," synchronized with the black bird logo, not the figure. * **D. Analysis of combat sequences:** This is discussed later in item 11 (‘[20.0s - 30.0s]‘) regarding "two separate battles."
Conclusion: The audio event synchronized with the moment the Obsidian Fury figure first appears is the speaker introducing himself and presenting the figure.
Final Answer: A ✓**Correct**

Figure 11: **Daily-Omni Example.** The model aligns visual and audio timestamps to identify the exact sound occurring at a visual onset.

World-Sense: Procedural Event Sorting

Question: What is the process followed by the security officer after discovering a firearm?
Choices: [A] Open and search the bag, alert authorities, and then call the supervisor, [B] Alert authorities, call the supervisor, and then open and search the bag, [C] Call the supervisor, alert authorities, and then open and search the bag, [D] Open and search the bag, call the supervisor, and then alert authorities

TAC-V Caption Output: *[8.7s - 12.1s] [visual] Two women approach... ⟨speech⟩She has a firearm in her bag⟨/speech⟩, supervisor gestures...[14.6s - 19.1s] [visual] Female agent holds phone... text ⟨speech⟩Now they're calling the police⟨/speech⟩...[27.3s - 30.2s] [visual] Text ⟨speech⟩It takes about 4 minutes for the officer to arrive⟨/speech⟩ appears...[38.7s - 39.9s] [visual] TSA agent in blue uniform opens beige handbag on tray...*

LLM Reasoning Trace: Based on the detailed audio-visual caption provided, here is the step-by-step reasoning to determine the correct sequence of events:
1. **Discovery of the Firearm:** * At **[8.7s - 12.1s]** (Item 7), the text '⟨text⟩She has a firearm in her bag⟨/text⟩' appears while a woman in a beige coat approaches. A supervisor gestures toward the machine. * At **[12.3s - 16.6s]** (Item 9), agents behind the machine discuss the situation, and a female agent points at the screen, confirming the discovery.
2. **First Action: Alerting Authorities/Calling Police:** * Immediately following the discovery, at **[14.0s - 14.8s]** (Item 10), we hear a '⟨sfx⟩Phone dial tone' as a female agent picks up the phone. * At **[14.6s - 19.1s]** (Item 11), on-screen text confirms: '⟨text⟩Now they're calling the police⟨/text⟩'. This establishes that contacting authorities/police is the immediate next step after identifying the threat on the screen.
3. **Delay/Waiting Period:** * At **[27.3s - 30.2s]** (Item 19), text appears stating: '⟨text⟩It takes about 4 minutes for the officer to arrive⟨/text⟩'. This indicates a waiting period for the authorities to arrive.
4. **Subsequent Actions (Opening/Searching):** * It is not until much later, at **[38.7s - 39.9s]** (Item 29), that we see a visual of a TSA agent opening a beige handbag (presumably belonging to the woman in the beige coat identified in Item 7) to remove items. * Prior to this, the bag remained on the belt or under observation while the call was made and the officers waited.
Sequence Analysis: The sequence observed is: 1. Firearm spotted on X-ray/Screen. 2. Agents discuss and Supervisor is involved (Item 7/9). 3. Agent calls the police (Alert authorities) (Item 11). 4. After the authorities arrive/time passes, the bag is physically opened and searched (Item 29).
Comparing this to the options: * A: Open and search -⟩ Alert -⟩ Call. (Incorrect order) * B: Alert authorities -⟩ Call supervisor -⟩ Open and search. (Close, but the "Call supervisor" happens almost simultaneously with discovery or just before the police call, but the distinct "Open and search" happens *last*). * C: Call supervisor -⟩ Alert authorities -⟩ Open and search. (This fits best. The supervisor is gesturing/involved at 8.7s immediately upon discovery. Then the phone call to police happens at 14.6s. Then the search happens at 38.7s). * D: Open and search -⟩ Call -⟩ Alert. (Incorrect order).
Let's refine the "Call supervisor" vs "Alert authorities" order. - Item 7 (8.7s): Supervisor gestures toward machine. (The supervisor is already alerted/present). - Item 11 (14.6s): Text "Now they're calling the police".
The flow is clearly: Identify/Supervisor involvement -⟩ Call Police -⟩ Search Bag. The bag is absolutely *not* opened first. Between options B and C, the distinction is subtle regarding the supervisor. However, usually, a line agent spots it, alerts a supervisor, and then authorities are called. The caption shows the supervisor gesturing *before* the phone call to police. Therefore, the supervisor is alerted/involved first, then authorities are alerted via phone, and finally, the bag is searched.
Therefore, the order is: Call/Involve Supervisor -⟩ Alert Authorities -⟩ Open and Search.
Final Answer: C ✓**Correct**

Figure 12: **World-Sense Example.** The model reconstructs a chronological protocol from the timestamped event log.

Prompt Template for MMAU & MMAR

```
You are an expert audio-understanding QA system.
You will be given:  (1) A multiple-choice question with answer
options (2) An "audio caption" describing detected sound events
with timestamps + a speech transcription
Your job:  - Use ONLY the information in the provided
caption/events/transcription.  - Pick exactly ONE option from the
choices.  - If the caption/transcription is insufficient to be
confident, still choose the most likely option, but lower your
confidence and explain why.  - Think step by step and provide
your answer.  - Return your answer in the following JSON format:
{"answer":  <answer>}
Now solve the following question:
{question} Answer options:  {choices} Caption with timestamps and
transcription:  {caption}
```

Figure 13: The standard prompt template used for the **MMAU** and **MMAR** benchmarks.

Prompt Template for MMSU

```
You are an expert audio-understanding QA system.
You will be given:  (1) A multiple-choice question with answer
options (2) An "audio caption" describing detected sound events
with timestamps + a speech transcription
Your job:  - Use ONLY the information in the provided
caption/events/transcription.  - Pick exactly ONE option from
the choices.  - If the caption/transcription is insufficient to
be confident, still choose the most likely option.  - Think step
by step and provide your answer.  - Return your answer in the
following JSON format:  {"answer":  <answer>}
Choose the most suitable answer from options A, B, C, and D to
respond the question in next line, **you should only choose A or B
or C or D.** Do not provide any additional explanations or content.
Question:  {question}
{choices}
Caption with timestamps and transcription:  {caption}
```

Figure 14: The prompt template used for the **MMSU** benchmark, which includes specific instruction tuning for option selection (A–D).

A. MMAU-Pro: Single-Audio Multiple Choice

```
You are an expert audio-understanding QA system. You will be given:
(1) A multiple-choice question with answer options (2) An "audio
caption" describing detected sound events... Your job: - Use
ONLY the information in the provided caption... - Pick exactly
ONE option from the choices. - Return your answer in the following
JSON format: {"answer": <answer>}
Choose the most suitable answer from options {letters} to
respond to the question. **You should only choose one letter
({letters}).**... Question: {question} {choices} Caption...:
{caption}
```

B. MMAU-Pro: Multi-Audio Multiple Choice

```
You are an expert audio-understanding QA system. You will be given:
(1) A multiple-choice question with answer options (2) Multiple
"audio captions" (labeled Audio 1, Audio 2, etc.) each describing
detected sound events... Your job: - Use ONLY the information
in the provided captions... from ALL audio files. - Consider
information from each audio file when answering the question. -
Pick exactly ONE option from the choices. - Return your answer in
the following JSON format: {"answer": <answer>}
Choose the most suitable answer from options {letters}... Question:
{question} {choices} Caption...: {caption}
```

C. MMAU-Pro: Single-Audio Open-Ended

```
You are an expert audio-understanding QA system. You will be
given: (1) An open-ended question about audio content (2) An
"audio caption" describing detected sound events... Your job: -
Use ONLY the information in the provided caption... - Provide a
concise, accurate answer based on the audio description. - Return
your answer in the following JSON format: {"answer": <answer>}
Question: {question} Caption...: {caption}
```

D. MMAU-Pro: Multi-Audio Open-Ended

```
You are an expert audio-understanding QA system. You will be given:
(1) An open-ended question about audio content (2) Multiple "audio
captions" (labeled Audio 1, Audio 2, etc.)... Your job: - Use
ONLY the information in the provided captions... from ALL audio
files. - Consider information from each audio file when answering
the question. - Provide a concise, accurate answer based on the
audio descriptions. - Return your answer in the following JSON
format: {"answer": <your answer>}
Question: {question} Caption...: {caption}
```

Figure 15: Prompt variations for the **MMAU-Pro** benchmark. We construct specific prompts depending on whether the task involves a single audio clip or multiple clips, and whether the output requires a multiple-choice selection or an open-ended response.

A. AVHBench: Audio-Visual Captioning

```
You are an expert audio-visual understanding QA system. You will
be given: (1) A request to describe the audio-visual content
(2) A detailed caption describing visual scenes, audio events,
and speech with timestamps Your job: - Use ONLY the information
in the provided caption (shot_list). - Describe what you see
and hear in a single coherent sentence. - Include both visual
elements (scenes, objects, actions) and audio elements (sounds,
speech, music). - Return your answer in the following JSON format:
{"answer": "<your description>"}
Describe what you see and hear based on the caption provided. Task
Type: AV Captioning Question: {question} Audio-Visual Caption
(shot_list with timestamps): {caption}
```

B. AVHBench: Audio-Visual Matching

```
... - Evaluate whether the audio content (sounds, speech, music)
is semantically consistent with the visual content (scenes, objects,
actions). - Consider: Do the sounds match what is visually
shown? Is the speech relevant to the visual context? - Answer
with exactly "Yes" or "No". ... Determine if the audio and visual
content are semantically matching. Answer with exactly "Yes" or
"No". Task Type: AV Matching Question: {question} Audio-Visual
Caption...: {caption}
```

C. AVHBench: Video-Driven Audio Hallucination

```
... (1) A question asking whether a specific object/entity is
making sound in the audio ... - Focus on the AUDIO descriptions:
[background], [sfx], [speech], [music] tags indicate audio content.
- Check if the specific object/entity mentioned in the question
is EXPLICITLY described as producing sound. - If the object
is only visible but not described as making sound, answer "No".
... Determine if the specified object/entity is making sound
in the audio. Answer with exactly "Yes" or "No". Task Type:
Video-driven Audio Hallucination Question: {question} Audio-Visual
Caption...: {caption}
```

D. AVHBench: Audio-Driven Video Hallucination

```
... (1) A question asking whether a specific object/entity is
visible in the video ... - Focus on the VISUAL descriptions:
[visual] tags and scene descriptions indicate visual content. -
Check if the specific object/entity mentioned in the question is
EXPLICITLY described as being visible. - If the object is only
heard but not described as visible, answer "No". ... Determine
if the specified object/entity is visible in the video. Answer
with exactly "Yes" or "No". Task Type: Audio-driven Video
Hallucination Question: {question} Audio-Visual Caption...:
{caption}
```

Figure 16: Prompt variations for **AVHBench**. We utilize specific instructions for hallucination detection to ensure the model distinguishes between what is seen (visual tags) and what is heard (audio tags).

Prompt Template for Video-Holmes

```
You are an expert video understanding and reasoning QA system.
You will be given:  (1) A reasoning question about video content
(may involve temporal, causal, or multi-hop reasoning) (2) A
detailed caption (shot_list) describing visual scenes, audio events,
and speech with timestamps
Your job:  - Use ONLY the information in the provided caption
(shot_list).  - Pay close attention to:  * Temporal sequences and
timing of events * Cause-and-effect relationships * Character
actions and their implications * Audio-visual synchronization
* Hidden meanings and implications from subtle details - Think
step by step and reason carefully through the question.  - Pick
exactly ONE option from the choices provided (A, B, C, D, E, or
F). - Return your answer in the following JSON format:  {"answer":
"<letter>"}
Analyze the video content carefully and choose the most suitable
answer.  **You should only respond with a single letter (A, B, C, D,
E, or F).** ...
Question Type:  {type_desc} Question:  {question} Options:
{options} Video Caption (shot_list with timestamps):  {caption}
```

Figure 17: The prompt template used for the **Video-Holmes** benchmark, emphasizing temporal and causal reasoning.

A. Prompt Template for Daily-Omni

```
You are an expert audio-visual understanding QA system.  ...
Your job:  - Use ONLY the information in the provided caption
(shot_list).  - Pay attention to the synchronization between audio
and visual elements.  - Pay attention to the visual and audio
information closely.  - Pick exactly ONE option from the choices
(A, B, C, or D). - Return your answer in the following JSON format:
{"answer":  <letter>}
Choose the most suitable answer from the given options.  **You
should only respond with A, B, C, or D.** ...  Question:  {question}
Options:  {options} Audio-Visual Caption (shot_list with
timestamps):  {caption}
```

B. Prompt Template for WorldSense

```
You are an expert audio-visual understanding QA system.  ...
Your job:  - Use ONLY the information in the provided caption
(shot_list).  - Pay attention to the synchronization between audio
and visual elements.  - Pay close attention to temporal information
(when events occur) and spatial relationships.  - Pick exactly ONE
option from the choices provided.  - Return your answer in the
following JSON format:  {"answer":  "<letter>"}
Choose the most suitable answer from the given options.  **You
should only respond with the letter (A, B, C, or D).** ...
Task Domain:  {task_domain} Task Type:  {task_type} Question:
{question} Options:  {options} Audio-Visual Caption (shot_list with
timestamps):  {caption}
```

Figure 18: Prompt templates for **Daily-Omni** and **WorldSense**, focusing on synchronization and spatial/temporal relationships.

System Prompt for VLM-based Captioning (TAC-V)

```
For each numbered audio event below, write a visual description
based on the video frames.
## Audio Events (with timestamps) {shot_list} Note:  Confidence
scores in parentheses indicate how certain the audio model was.
Low confidence (< 0.7) suggests uncertainty--prioritize correcting
these when visuals disagree.
## Step 1:  Reasoning First, inside <reasoning> tags, analyze:  -
What is the overall scene/setting?  - For each numbered event,
look at frames near the START of the timestamp range.  - Note
any audio labels that need fixing based on visuals (fix silently
in descriptions).  - Pay attention to confidence scores:  low
confidence sounds are good candidates for correction if visuals
suggest something different.
## Step 2:  Descriptions Then, inside <descriptions> tags,
write exactly {num_entries} lines:  <reasoning> Scene is a busy
restaurant kitchen...  Event 2 (2.0-4.5s):  Sizzling sound matches
pan on stovetop...  </reasoning> <descriptions> [ 0.0s - 3.0s]
[visual] Industrial kitchen with stainless steel counters...
[ 2.0s - 4.5s] [sfx] Pan sizzles on gas burner as chef flips
vegetables...  [ 3.5s - 6.0s] [speech] Head chef in tall white hat,
lips moving...  </descriptions>
## Rules - FUSE audio + video:  keep audio description AND add
visual context.  - CRITICAL: Fix audio labels that don't match
visuals.  The audio model confuses acoustically similar sounds
(e.g., "helicopter" → "fan whirs", "applause" → "rain patters").
- USE VARIED SOUND DESIGN VOCABULARY for [sfx] (e.g., impact,
whoosh, drone, riser).  - [speech]:  Describe who is speaking
and how they sound, but NEVER the content.  - [sfx]:  Keep the
CORRECTED sound label + add visual source.  - Describe what you
SEE, not emotions or inner states.
```

Figure 19: **VLM System Prompt.** The prompt enforces a "Reason-then-Describe" workflow, explicitly instructing the model to use visual evidence to correct low-confidence audio predictions (hallucinations) before generating the final dense captions.

Prompt Template for Speaker Diarization

```
Identify each speaker's turns.  Give start and end times for each
turn.  Output format -- anonymous Speaker N (1-indexed by first
appearance in the 30-second window), one event per turn, multi-turn
speakers get comma-separated time ranges.  Empty windows return 0
speakers.
```

Figure 20: The prompt template used for speaker diarization during both training and inference (Section 4.2).

Speaker Diarization: Sample Outputs

Example 1: Two-speaker conversation
Speaker 1 from 1.95s to 9.29s. Speaker 2 from 11.51s to 15.00s.

Example 2: Three speakers with multi-turn
Speaker 1 from 0.00s to 2.35s, 5.16s to 6.36s, 11.39s to 14.13s. Speaker 2 from 2.43s to 3.58s. Speaker 3 from 8.53s to 9.08s.

Example 3: Single speaker
Speaker 1 from 0.00s to 11.63s.

Example 4: Silent window
0 speakers.

Figure 21: **Speaker Diarization Examples.** Sample TAC outputs on 30-second audio windows, showing multi-speaker turn detection with precise timestamps, multi-turn tracking via comma-separated ranges, and correct handling of silent windows.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction (Section 1) clearly state the three contributions—TAC, TAC-V, and the describe-then-reason cascade—and the experimental results in Sections 4.1–5.2 support each claim.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses limitations including over-estimation of dramatic events due to synthetic data bias and lack of fine-grained musical precision (e.g., chord progressions).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results or proofs. It is an empirical systems paper.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4 details the training setup (hardware, batch size, LoRA rank, learning rate, optimizer, iterations). Section 3 describes the full data pipeline and Algorithm 1 formalizes the mixing procedure. Evaluation prompts are provided in Appendix H.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The synthetic data pipeline uses proprietary licensed audio sources. We provide a detailed algorithmic description (Algorithm 1) and training details sufficient for reproduction with alternative audio sources.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 specifies GPU type, batch size, LoRA rank/alpha, optimizer (AdamW), learning rate schedule, warmup steps, and number of training iterations. Evaluation benchmarks and their standard protocols are cited.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because all training runs use the same seed and data for controlled ablation. The computational cost of multiple full training runs is prohibitive. We mitigate this by reporting comprehensive ablation studies that demonstrate robustness across hyperparameter choices.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4 states training is performed on 8 NVIDIA A100 (80GB) GPUs with 5k iterations and a batch size of 32.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics. We discuss potential societal impacts, including surveillance risks and bias considerations, in Section C.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section C discusses positive impacts (accessibility tools, safety-critical monitoring) and negative impacts (surveillance risks, potential bias inherited from source libraries).

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The model produces descriptive text captions of audio/video content and does not generate audio, video, or other media. It poses low risk for direct misuse compared to generative models.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets (AudioCaps, Clotho, AudioSet-Strong, TACOS, MMAU, etc.) and models (Qwen2-Audio, Whisper, FLAM, Qwen3-VL) are properly cited. The source audio for synthetic mixing is from licensed libraries.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets or model checkpoints at this time. The synthetic data pipeline is described algorithmically in Algorithm 1.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are a core component of our methodology. We use GPT-OSS-120B and Qwen-32B-VL for generating pretraining captions (Section 3), Gemini 3 Pro and Qwen3-Thinking as downstream reasoners (Section 5), and Qwen3-VL-32B as the VLM in TAC-V (Section 3.4). All model names and versions are disclosed. See also Appendix I.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.