Auto-AEG: Scalable Data Construction for Open-Vocabulary Audio Event Grounding

Zihan Zhang[1,*], Xize Cheng[1,*], Wenhao Yan[2,*], Tong Zhang[1], Dongjie Fu[1], Boyun Zhang[1], Yongbo He[1], Tao Jin[1,†]

[1]Zhejiang University, [2]Tsinghua University

jint_zju@zju.edu.cn

[*]Equal contribution. [†]Corresponding author.

Abstract

Large Audio-Language Models (LALMs) reason fluently about sound yet struggle to localize precisely *when* events occur, while classical Sound Event Detection attains frame-level precision only over a closed label set. At the intersection of these paradigms lies the task of *Open-Vocabulary Audio Event Grounding*: predicting all time intervals of a target sound event described by an arbitrary natural language query. While this task is crucial for real-world audio understanding and LALM adaptation, it is bottlenecked by data scarcity. Few large-scale resources provide open-vocabulary onset/offset supervision, and manual temporal annotation is prohibitively expensive.

To address this, we introduce **Auto-AEG**, a scalable pipeline that constructs such supervision by automatic data construction and model fine-tuning. It pairs programmatically synthesized clips, which carry exact ground-truth intervals for supervised cold-start, with multi-model pseudo-labels on real-world audio that supply the reward signal for reinforcement learning. Training with this pipeline yields promising performance gains on both the DESED SED benchmark and **AEGBench**, an independent difficulty-stratified benchmark we release. Our results show that automatically constructed data, coupled with interval-aware reward function design, is an effective data-side route to expanding the temporal localization capability of LALMs.

Keywords: audio event grounding, large audio-language models, temporal localization, data construction, reinforcement learning, sound event detection

1. Introduction

Large Audio Language Models (LALMs) [7, 12, 28, 4, 10] have demonstrated remarkable capabilities in audio understanding and reasoning, bridging the acoustic and linguistic domains by aligning audio encoders with large language models. As these models grow increasingly proficient at describing, classifying, and reasoning about sound content, the ability to precisely localize sound events in time emerges as a critical yet underexplored frontier.

Classical Sound Event Detection (SED) systems achieve frame-level temporal precision yet remain constrained by closed-label vocabularies that cannot generalize to the richness of natural language, while LALMs handle arbitrary queries with remarkable flexibility but struggle to produce fine-grained temporal predictions. We study the task that sits between these paradigms, which we call Open-Vocabulary Audio Event Grounding: given an audio clip and a natural language query describing a target sound event, predict all onset/offset intervals where that event occurs, supporting multiple occurrences and an open event vocabulary.

Progress on this field is fundamentally limited by data scarcity. Existing temporal audio grounding datasets [37, 21] are small, narrow in label diversity, and the cost of manually annotating fine-grained boundaries for arbitrary sound events is prohibitive. The field therefore lacks a viable training resource even though the task is within reach of current LALMs. Our central claim is that this gap can be closed on the *data* side, without architectural change, by constructing supervision automatically and exploiting it with reinforcement learning.

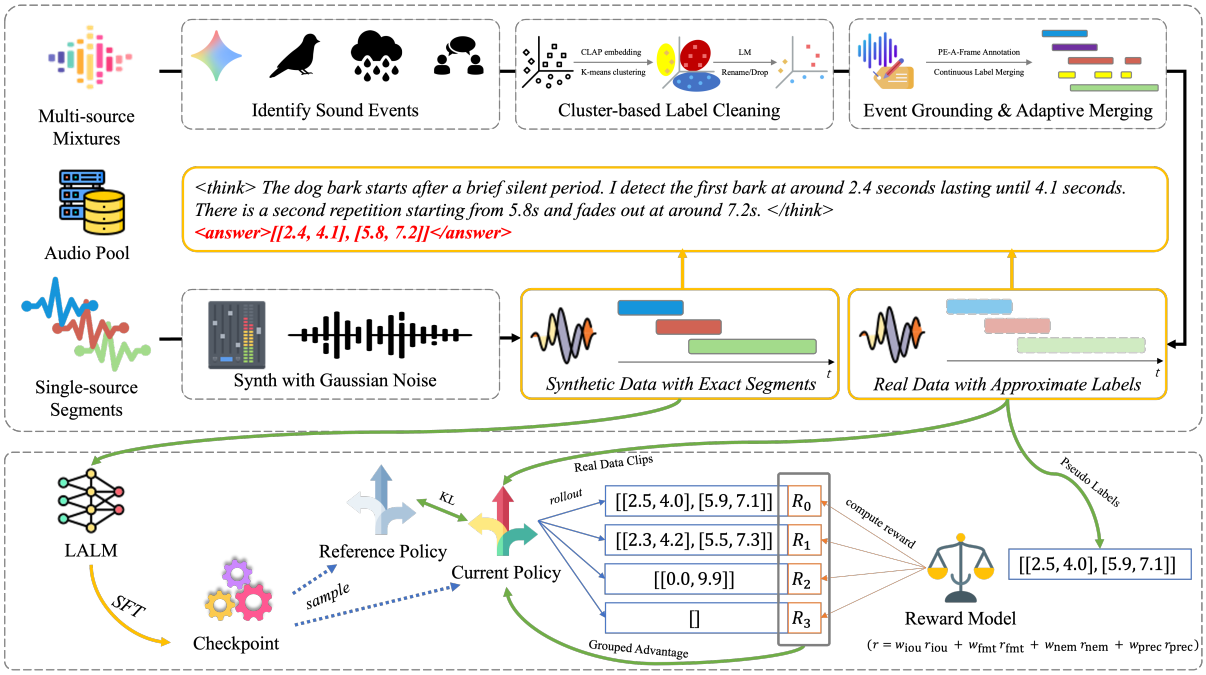

Figure 1: Auto-AEG pipeline overview. The pipeline annotates real-world audio with pseudo-labels for GRPO, while programmatically synthesizing clips with exact ground-truth intervals for SFT cold-start.

We instantiate this with **Auto-AEG** (Figure 1), a scalable pipeline whose key insight is that the two kinds of data a grounding policy needs have complementary, annotation-free acquisition strategies. *Synthetic* clips, composed programmatically from real audio, carry exact ground-truth intervals and are ideal for an SFT cold-start; *real-world* clips, pseudo-labeled through multi-model collaboration, are imperfect but provide the informative reward signal exploited by Group Relative Policy Optimization (GRPO). Both stages scale without manual annotation: synthetic generation is unbounded, and pseudo-labeling extends to any public audio collection. We pair this with **AEGBench**, an independent, quality-filtered evaluation benchmark drawn from four diverse audio sources and stratified by hard-case difficulty for diagnostic analysis.

Training Qwen3-Omni-30B-A3B and Qwen2.5-Omni-7B on Auto-AEG data with a two-stage fine-tuning approach that contains interval-aware reward design substantially improves temporal localization on AEGBench, with mIoU gains of **+73.9%** and **+23.1%** over zero-shot baselines and subtle degradation of general audio understanding. These results confirm Auto-AEG as an effective data-side solution for scaling LALM temporal capability.

Our contributions can be summarized as follows:

(i) *Auto-AEG*, a scalable pipeline that automatically constructs open-vocabulary onset/offset supervision, providing both exact-GT synthetic data and in-the-wild pseudo-labels;

(ii) *AEGBench*, an independent quality-filtered benchmark with difficulty stratification;

(iii) a two-stage fine-tuning framework that couples an instruction-tuned cold start with interval-aware GRPO, which turns imperfect pseudo-labels into large, architecture-free gains in LALM temporal grounding.

2. Related Work

2.1. Video Temporal Grounding

Temporal sentence grounding in video (TSGV) has been studied extensively since Gao et al. [9] and Anne Hendricks et al. [2]. Methods range from proposal-based approaches [39] to proposal-free span prediction [40] and transformer-based moment retrieval [17, 20]. Large video-language models have more recently addressed TSGV in a generative setting, with Time-R1 [18] showing that explicit temporal supervision can substantially improve timestamp reasoning in video LLMs—a result that directly motivates our data-construction approach for the audio modality.

2.2. Sound Event Detection

Sound event detection (SED) is the immediate predecessor of our task. SED has been studied extensively through the DCASE challenge series [13], with dominant architectures including CRNNs [41], PANNs [16], and self-supervised transformers such as AST [11] and BEATs [3]. Weakly supervised

SED [25, 30] uses clip-level labels to avoid expensive frame-level annotation—a design philosophy aligned with our pipeline, which leverages pretrained model outputs rather than manual temporal annotation.

2.3. Audio Temporal Grounding and Moment Retrieval

Text-to-Audio Grounding (TAG) [37] introduced the AudioGrounding dataset (4,662 clips), pairing onset/offset timestamps with phrase-level queries. Xu et al. [38] extend this to weakly supervised regimes. Audio Moment Retrieval (AMR) [21] scales temporal grounding to longer untrimmed audio. CASTELLA [22] provides the first real-world AMR benchmark with 1,862 recordings of 1–5 minutes. BLAB [1] evaluates long-form localization focusing on speech. TACOS [23] contributes 12,358 Freesound recordings with human-verified temporal boundaries and trains a frame-level contrastive model; it is the closest concurrent work but targets contrastive retrieval rather than LALM adaptation. TimeAudio [32] improves temporal reasoning through architectural modifications trained on a 260K-sample supervised dataset.

SpotSound [27] adapts LALMs to open-vocabulary temporal grounding and augments existing grounding corpora with synthetic clips formed by mixing foreground events into background ambience. The idea of synthesizing temporal supervision by composition is itself established—Clotho-Moment [21], AudioTime [34], and PicoAudio [35] all generate timestamp supervision automatically.

Our work innovates in two respects compared to the above concurrent works. First, in *training paradigm*: SpotSound relies on supervised fine-tuning with a negative-sampling objective, whereas we exploit the imperfect real-world pseudo-labels as a reward signal and show that GRPO is what converts automatically constructed data into large temporal-grounding gains. Second, in *scope*: we target large Omni models (up to 30B) and pair the pipeline with an independent, difficulty-stratified benchmark rather than reusing existing grounding test sets. To our knowledge, Auto-AEG is the first to couple annotation-free data construction with reinforcement learning for open-vocabulary audio event grounding.

These resources rely on human annotation or supervised training and remain at the scale of thousands to tens of thousands of clips. Auto-AEG instead fills the training data gap by *automatically constructing* onset/offset supervision for both LALM SFT and GRPO from existing public audio.

2.4. Audio Language Models

LALMs [28, 6, 5, 14] have achieved strong performance on auditory understanding tasks. Qwen2-Audio [5], Qwen2.5-Omni [29], and Qwen3-Omni [36] are Whisper-based models that process audio as a sequence of continuous encoder features and generate responses autoregressively. Despite their strong semantic competence, these models lack targeted temporal boundary supervision—the gap our data directly addresses.

2.5. Automated Data Construction and RL Fine-Tuning

Large-scale automated annotation has been used in vision and audio: pseudo-label pipelines using CLIP [24] generate bounding boxes for downstream training; WavCaps [19] used ChatGPT to clean captions at scale. Reinforcement learning via GRPO [26] has been successfully applied to improve reasoning in language models [18], including temporal understanding in video. We bring this paradigm to open-vocabulary temporal audio grounding, using automatically constructed pseudo-labels as the reward signal rather than curated supervision.

3. Task Definition

Formal Definition. Let $\mathbf{x} \in \mathbb{R}^T$ be an audio waveform of duration D seconds and q a natural language query describing a target sound event type. The model must predict a set of time intervals $\hat{\mathcal{Y}} = \{[s_i, e_i]\}_{i=1}^{K}$, where each interval corresponds to an occurrence of the event described by q in $\mathbf{x}$. Ground-truth segments are denoted $\mathcal{Y}^* = \{[s_j^*, e_j^*]\}_{j=1}^{M}$. This formulation requires the model to: (a) determine how many occurrences exist without count supervision, (b) handle zero occurrences, and (c) generalize over open event vocabulary.

Response Format. Models are prompted to list all matching intervals in JSON array format within `<answer>` tags, optionally preceded by a `<think>` reasoning chain:

```
<think> ... </think>
<answer> [[s1, e1], [s2, e2], ...]
↪ </answer>
```

An empty list `[]` indicates the target event is absent. All metrics are computed from the intervals extracted from the `<answer>` block.

Evaluation Metrics. We use the following metrics:
mIoU: Mean IoU between each ground-truth interval and its best-matching prediction, averaged over

present-event queries ($|\mathcal{Y}^*| > 0$):

$$\text{mIoU} = \frac{1}{|\mathcal{Y}^*|} \sum_j \max_i \text{IoU}\big([s_j^*, e_j^*], [s_i, e_i]\big) . \quad (1)$$

This is a *recall-oriented* localization score: each ground-truth segment is credited for the prediction that best overlaps it, and extra or hallucinated intervals are not penalized. To control for over-prediction we additionally report *Event F1* and *onset Precision*, which use symmetric matching and explicitly penalize spurious predictions; the AEG-Bench improvements reported in Section 7.1 hold under all three. Absent-event queries ($|\mathcal{Y}^*| = 0$, whose correct answer is the empty list `[]`) are excluded from mIoU; their correct rejection is instead measured by Event F1 and the precision metrics, where any non-empty prediction counts as a false positive.

Recall-IoU@θ (R@θ): Fraction of ground-truth segments matched by at least one prediction at IoU $\geq \theta$, at $\theta \in \{0.3, 0.5, 0.7\}$.

Precision-IoU@θ: the dual quantity—the fraction of *predicted* segments matched by at least one ground-truth interval at IoU $\geq \theta$. Their harmonic mean, **F1-IoU@θ**, is used as the localization component of the GRPO reward (Section 6.2).

Event F1 (ev_F1): Standard F1 computed by matching predicted and ground-truth intervals at IoU ≥ 0.5 under one-to-one assignment, yielding explicit true-/false-positive counts. F1-IoU and Event F1 track the same trend but differ in matching: F1-IoU permits many-to-one overlap matching and is smoother as a per-sample training reward, whereas Event F1 enforces one-to-one matching and is stricter as an evaluation metric.

Segment F1 (seg_F1) and **Onset P/R**: Frame-level and onset-level precision/recall following DCASE SED evaluation convention.

Difficulty Taxonomy. We identify six hard-case categories in the audio modality, as shown in Table 1.

Category	Description
Polyphonic overlap (PO)	Two or more distinct events co-occur; the query targets one.
Gradual onset/offset (GO)	The event begins or ends gradually (>500 ms), making precise boundary placement ambiguous.
Repeated occurrence (RO)	The same event type recurs multiple times; the model must emit all instances.
Low-contrast event (LC)	The event is only moderately louder than its background (12–28 dB contrast), making the boundary harder to localize precisely.
Semantic ambiguity (SA)	The query description does not uniquely determine the target interval.
Long duration (LD)	The target event spans more than 30 seconds.

Table 1: Hard-case categories in AEGBench.

4. Auto-AEG: Scalable Data Construction

Constructing AEG supervision faces a fundamental tension: synthetic data can be assigned exact ground-truth intervals by construction, but it diverges from the acoustics of natural recordings; real-world audio better matches deployment conditions, but automatic temporal annotations are inherently noisy. Auto-AEG resolves this by matching each data type to the training objective it serves. Stage 1 generates synthetic data with *exact* ground truth—suitable for SFT, which optimizes directly against the target sequence and is therefore sensitive to label noise; Stage 2 annotates real-world audio with *pseudo-labels* suited to the noise-tolerant GRPO objective, which learns from a scalar reward rather than token-level supervision and can self-correct over multiple rollouts even when individual annotations contain errors.

4.1. Audio Pool Construction

Both stages draw from a pool of audio clips sourced from the FreeSound subset of WavCaps [19], chosen for its broad event-label diversity, permissive licensing, and community-supplied tags that serve as initial label seeds. Each clip is pre-screened by PE A-Frame [31] in frame-level scoring mode: clips where the stated tag produces no active frames above threshold are discarded before any further annotation. This early gate eliminates clips whose tag is acoustically absent or too quiet to localize reliably—retaining them would introduce systematically wrong pseudo-labels in Stage 2 with no downstream signal indicating the annotation is invalid. Each retained entry stores its audio, tag label, and detected intervals, and serves as the common source for both data stages.

4.2. Stage 1: Programmatic Synthesis for SFT Cold-Start

No large-scale dataset provides open-vocabulary onset/offset annotations, so Stage 1 generates

10,000 training clips by programmatically composing pool segments. For a randomly chosen target label, we lay down several non-overlapping occurrences, optionally add distractor events of other labels, and mix the result onto a Gaussian background at a controlled signal-to-noise ratio. The occurrence count is deliberately skewed toward multiple events per clip: a single-occurrence majority would bias the model to always emit exactly one interval and systematically under-detect multi-event queries at inference time. Because every interval is fixed by the composition script, the ground truth $\mathcal{Y}^* = \{[s_j, e_j]\}$ is exact and noiseless.

This clean signal provides a stable cold-start that instills the response format and basic acoustic recognition before the model faces real-world variability. Each clip is paired with a List-All query asking for every interval of the target event as a JSON array. Detailed description of sampled recordings, sampling ranges and the exact prompt are given in Appendix B.

4.3. Stage 2: Multi-Model Annotation for GRPO

Stage 2 annotates real FreeSound clips (10–30 s) to produce training data for GRPO. The annotation pipeline decouples *what* events are present from *when* they occur, because no single model handles both reliably: LALMs excel at semantic event identification but produce coarse or hallucinated temporal boundaries, while dedicated frame-level models locate boundaries precisely but require a text query as input.

Gemini identifies audible events from fixed-length 10-second chunks—chunking improves label focus on long clips where full-clip prompting dilutes attention across many events—and pools labels across chunks into a compact, semantically orthogonal inventory per clip. Each label is then classified as *continuous* (sustained events such as engine noise or rain) or *discrete* (transient events such as a bark or knock) before temporal localization. This distinction controls span-merging behavior: without it, a uniform rule would incorrectly fuse two consecutive bark events separated by a brief silence, or fail to bridge legitimate millisecond dropouts within a continuous rain shower—both are common labeling errors that degrade boundary quality. PE A-Frame [31] then localizes each label by thresholding per-frame audio–text similarity into onset/offset spans, merging adjacent spans for continuous-type labels only.

Because labels are coined independently per clip, near-synonyms accumulate across the dataset (e.g., *car horn*, *vehicle horn*, *horn honking*), fragmenting the training vocabulary and making the same physical event appear under multiple query strings. A final global pass addresses this: all unique labels are embedded with CLAP [33] and clustered by acoustic similarity, then a language model assigns *keep*, *drop*, or *rename* to each label within its cluster. Clustering groups acoustically similar candidates that embedding distance alone cannot distinguish in quality, and the LM applies specificity and redundancy judgements that a nearest-neighbor rule cannot. Prompts, thresholds, and merging rules are detailed in Appendix C.

Table 2 summarizes both data stages.

Stage	**Clips**	**Queries**	**GT Type**
Stage 1 (Synthetic SFT)	10,000	10,000	Exact
Stage 2 (Real GRPO)	2,000	5,244	Pseudo
Total	12,000	15,244	—

Table 2: Auto-AEG dataset statistics. Stage 1 synthetic data has exact programmatically-determined ground truth; Stage 2 real data has pseudo-labels from multi-model annotation.

5. AEGBench: Difficulty-Stratified Benchmark

AEGBench provides an independent, quality-controlled evaluation surface for open-vocabulary audio event grounding. It is constructed entirely separately from Auto-AEG training data to avoid circular evaluation.

5.1. Candidate Selection and Quality Filtering

Candidates are drawn from four complementary sources—AudioSet Strong Labels, the FSD50K eval split [8], the BBC Sound Effects Library, and short YouTube life-sound clips—spanning human-verified strong labels, community labels, and professional field recordings. Each candidate must pass an **energy-contrast** filter requiring the target event to be measurably louder than its background, together with active-ratio, duration, and per-category-cap constraints (Appendix D). The energy-contrast criterion is the primary differentiator from prior benchmark construction, which relies on duration and silence-ratio filters: it directly quantifies whether the signal would yield a reliable, unambiguous temporal boundary, making it a more

principled quality gate. Per-source counts after filtering are reported in Table 3.

5.2. Annotation and Hard-Case Tagging

Benchmark items are annotated by the same multi-model pipeline as Stage 2 of Auto-AEG—Gemini label identification, event-type classification, PE A-Frame localization, and CLAP-based global label cleaning—yielding clean onset/offset annotations over a canonical vocabulary. Each item is then tagged with the applicable difficulty categories from the taxonomy of Section 3 (polyphonic overlap, gradual onset/offset, repeated occurrence, low-contrast, long duration, and semantic ambiguity) to enable stratified diagnostic evaluation. The tagging rules and thresholds are given in Appendix D.

5.3. Human Verification

After automated annotation, all 3,427 candidate items undergo human review. An annotator examines each item and: (1) confirms or corrects the identified sound label; (2) adjusts onset/offset boundaries to match their auditory perception where the automatic annotation is incorrect. This review results in timestamp corrections and deletions of invalid items, producing a final benchmark of **3,427 items**. All benchmark items are kept strictly disjoint from the Auto-AEG training splits.

5.4. Benchmark Statistics

Source	Items	Fraction
AudioSet Strong Labels	2,230	65.1%
FSD50K eval	954	27.8%
BBC Sound Effects	234	6.8%
YouTube Life Sounds	9	0.3%
Total	3,427	100%

Table 3: AEGBench source distribution (after human verification and deduplication). Energy contrast $\geq$12 dB and active-ratio filtering applied uniformly across all sources.

Table 3 shows the source distribution. Audio clips range from 10–120 s, with target event durations from 0.5–60 s. The six difficulty categories are not mutually exclusive; items belonging to the Polyphonic Overlap category form the largest difficulty subset.

5.5. Comparison with Existing Benchmarks

Table 4 positions AEGBench against existing temporal audio grounding resources across five key dimensions. Prior benchmarks such as AudioGrounding [37] and AMR [21] support natural-language queries but cover only short clips and provide no difficulty stratification. CASTELLA [22] extends to long-form recordings yet lacks both hard-case annotations and an associated training split. TACOS [23] is the only prior resource to pair an annotated test set with a large-scale training corpus, but its evaluation targets contrastive retrieval rather than open-vocabulary event localization, and its benchmark clips remain under 30 seconds. AEGBench is the only benchmark to simultaneously support natural-language queries, long-audio clips, difficulty-stratified hard cases, real-world recording sources, and a paired Auto-AEG training split covering the same open label vocabulary.

Benchmark	NL Query	Long Audio	Hard Cases	Real-World	Training Data
DCASE Task 4 (SED) [13]	✗	✗	✗	✓	✗
AudioGrounding [37]	✓	✗	✗	✗	✗
AMR [21]	✓	✗	✗	✗	✗
CASTELLA [22]	✓	✓	✗	✓	✗
TACOS [23]	✓	✗	✗	✓	✓
AEGBench (ours)	✓	✓	✓	✓	✓

Table 4: Comparison with existing audio temporal grounding resources. "Training Data" indicates a large-scale Auto-AEG training split is available alongside the benchmark; "Long Audio" indicates support for clips longer than 30 seconds.

6. Two-Stage Fine-tuning Framework

We use the Auto-AEG data to fine-tune LALMs in two sequential stages. The goal is not to introduce a new training algorithm, but to *validate data quality*: if our automatically constructed data improves temporal localization on the independently-annotated AEGBench, the pipeline has produced useful training signal. Both stages share a single prompt template (Figure 2, Appendix A) that elicits an explicit `<think>` reasoning trace before the final `<answer>`.

6.1. Stage 1: SFT Cold-Start

Backbones. We train Qwen3-Omni-30B-A3B-Thinking (hereafter Q3-Omni) and Qwen2.5-Omni-7B [29] (hereafter Q2.5-Omni) using QLoRA 4-bit NF4, compute dtype bfloat16, LoRA rank r=16, α=32, applied to $\{q, k, v, o, \text{gate}, \text{up}, \text{down}\}$_proj in the language-model component. The audio tower is kept frozen and cast to bfloat16.

Data and Format. SFT is performed on the 10,000 synthetic clips from Stage 1 of Auto-AEG, split 9:1 into train/val by random seed. Each example uses the List-All query format, with a `<think>` scaffold that states the queried event, the total clip duration, and enumerates each ground-truth occurrence by its onset/offset interval—exposing the temporal structure of each GT annotation in natural language and stabilizing training on the clean synthetic data.

Training Details. 3 epochs, learning rate 2×10^{-4}, batch size 2, gradient accumulation to effective batch 16. Best checkpoint selected by validation loss.

6.2. Stage 2: GRPO Fine-Tuning

Starting from the Stage 1 SFT checkpoint, we apply GRPO [26] on the 5,244 real-world queries from Stage 2 of Auto-AEG.

Reward Function. The total reward combines four components:

$$r = 0.65\, r_{\text{iou}} + 0.15\, r_{\text{fmt}} + 0.05\, r_{\text{nem}} + 0.15\, r_{\text{prec}}, \quad (2)$$

where:

- r_{iou}: **F1-IoU@0.5**—the harmonic mean of Recall-IoU@0.5 and Precision-IoU@0.5. Using F1 rather than pure recall simultaneously penalizes missed detections and hallucinated intervals, preventing the degenerate "dense-fill" strategy.
- r_{fmt}: **Format reward**—presence of `<think>` (+0.3) + `<answer>` (+0.3) + valid JSON array (+0.4).
- r_{nem}: **Non-empty reward**—+1 if the prediction contains at least one interval, lightly penalizing over-rejection.
- r_{prec}: **Precision penalty**—linearly decays when the ratio of predicted to ground-truth interval count exceeds 2; reaches zero at ratio = 4. This explicitly suppresses the tendency to pad outputs with spurious short intervals to inflate recall.

Because Stage 2 retains only labels that PE A-Frame successfully localizes (Section 4.3), every GRPO training query has at least one ground-truth interval ($|\mathcal{Y}^*| \geq 1$). The count-based terms r_{nem} and r_{prec} are therefore always well-defined, and we use the no-rejection query variant (Appendix A). The policy's ability to reject absent events is consequently *evaluated* on AEGBench (via Event F1 and precision) rather than trained as an explicit reward objective. GRPO advantage is computed as $\text{adv}_i = (r_i - \bar{r})/(\text{std}(r) + \varepsilon)$, with KL-divergence regularization to the frozen SFT reference policy (coefficient β=0.04). Training runs for 3 epochs with DeepSpeed ZeRO-2 on 2 GPUs, learning rate 5×10^{-5}, G=4 rollouts per sample.

6.3. External Zero-Shot Baselines

We evaluate four external zero-shot baselines to anchor the trained models against the broader landscape: **Gemini-3-Pro** (closed-source, accessed via an OpenAI-compatible multimodal endpoint with base64-encoded audio), **Kimi-Audio-7B-Instruct** [15] (same-scale open-source audio LLM with native temporal-grounding support), **Qwen2-Audio-7B-Instruct** [5] (a same-family predecessor of our Qwen-Omni backbones), and **Audio Flamingo Next** [10] (the next-generation open-source audio LLM in the Audio Flamingo series). All baselines receive the same List-All prompt and are not fine-tuned. We note that Audio Flamingo Next rarely emits intervals under this prompt (onset recall 0.046); its high onset precision (0.919) therefore reflects a small number of high-confidence predictions rather than strong coverage, so we report it for completeness and do not draw strong conclusions from its precision.

7. Experiments

7.1. Main Results on AEGBench

Table 5 reports open-vocabulary AEG results on the full AEGBench (n=3,427 items, 5,275 queries). All SFT+GRPO models use the Auto-AEG pipeline data only; no AEGBench items appear in training.

Model	mIoU	ev_F1	seg_F1	onset_P	onset_R	R@0.5
External zero-shot baselines						
Gemini-3-Pro	0.323	0.282	0.574	0.413	0.320	0.289
Kimi-Audio-7B	0.117	0.160	0.280	0.193	0.314	0.070
Qwen2-Audio-7B	0.157	0.187	0.348	0.538	0.281	0.083
Audio Flamingo Next	0.028	0.027	0.064	0.919	0.046	0.020
Auto-AEG SFT + GRPO						
Qwen3-Omni-30B	0.276	0.295	0.528	0.463	0.333	0.243
+ SFT	0.371	0.343	0.642	0.368	0.387	0.318
+ SFT + GRPO	**0.480**	**0.524**	**0.697**	**0.559**	**0.566**	**0.465**
Qwen2.5-Omni-7B	0.324	0.340	0.602	0.331	0.490	0.264
+ SFT	**0.424**	0.416	**0.657**	0.411	**0.508**	**0.402**
+ SFT + GRPO	0.399	**0.474**	0.609	**0.594**	0.435	0.384

Table 5: Main results on AEGBench (3,427 items, 5,275 queries). Bold marks the best value per model family and column.

Model	PO	GO	RO	LC	SA	LD
External zero-shot baselines						
Gemini-3-Pro	0.294	0.327	0.256	0.331	0.268	0.194
Kimi-Audio	0.114	0.130	0.115	0.108	0.115	0.066
Qwen2-Audio	0.145	0.167	0.137	0.153	0.140	0.028
Audio Flamingo Next	0.026	0.029	0.023	0.023	0.028	0.004
Auto-AEG SFT + GRPO						
Qwen3-Omni-30B	0.249	0.284	0.208	0.284	0.223	0.190
+ SFT	0.346	0.387	0.298	0.337	0.341	0.157
+ SFT + GRPO	**0.445**	**0.491**	**0.389**	**0.482**	**0.405**	**0.277**
Qwen2.5-Omni-7B	0.301	0.330	0.265	0.323	0.281	0.170
+ SFT	**0.403**	**0.439**	**0.355**	0.379	**0.400**	0.115
+ SFT + GRPO	0.357	0.403	0.275	**0.403**	0.309	**0.239**

Table 6: Per-category mIoU on AEGBench difficulty subsets. PO: polyphonic overlap; GO: gradual onset/offset; RO: repeated occurrence; LC: low-contrast; SA: semantic ambiguity; LD: long duration.

SFT Contribution. Synthetic-data SFT alone already delivers substantial gains: +34.4% relative mIoU for Q3-Omni (0.276 → 0.371) and +30.9% for Q2.5-Omni (0.324 → 0.424). This confirms that the programmatic synthesis pipeline produces learnable temporal grounding signal even without any real-world annotation.

GRPO Contribution. The effect of Stage 2 GRPO differs markedly between the two model families. For Q3-Omni, GRPO delivers consistent across-the-board improvements: mIoU rises from 0.371 to **0.480** (+29.4% relative), ev_F1 from 0.343 to 0.524, and onset precision from 0.368 to 0.559—the model learns to commit to genuine boundaries while simultaneously improving recall. For Q2.5-Omni, the picture is more nuanced: overall mIoU slightly decreases (0.424 → 0.399), yet event-level F1 climbs from 0.416 to 0.474 and onset precision rises substantially from 0.411 to 0.594. This divergence suggests that GRPO pushes the smaller 7B model toward a high-precision, lower-recall regime, trading overall interval coverage for sharper boundary estimates. The precision component of the reward is the likely driver: Q2.5-Omni responds by suppressing over-prediction more aggressively than Q3-Omni, at a cost to recall.

Cross-Generation Analysis. Q3-Omni (30B) achieves the best overall result (mIoU = 0.480 at SFT+GRPO), while Q2.5-Omni (7B) peaks earlier at the SFT stage (mIoU = 0.424). The fact that GRPO improves Q3-Omni on all metrics but produces a precision–recall trade-off for Q2.5-Omni suggests that the reward signal from our noisy real-world annotations is more readily exploited by the larger model, whose stronger priors allow it to improve boundary precision without sacrificing recall.

7.2. Hard-Case Analysis

Table 6 breaks down performance by the six difficulty categories defined in Section 5. Long Duration (LD) is the hardest category across *all*

models: even Gemini-3-Pro scores only 0.194 on LD compared to 0.327 on GO, and the best trained model (Q3-Omni SFT+GRPO) reaches only 0.277—consistent with the 30 s encoder window limit imposing a hard ceiling on long-event localization. Gradual Onset/Offset (GO) and Low-Contrast (LC) are comparatively tractable after training, with Q3-Omni SFT+GRPO reaching 0.491 and 0.482 respectively, both surpassing Gemini-3-Pro on these categories. The precision–recall asymmetry observed in the main results recurs here: for Q2.5-Omni, GRPO improves LC and LD (where boundary sharpness matters most) but slightly hurts PO, GO, RO, and SA relative to SFT.

7.3. Evaluation on SED benchmark

Evaluating whether the grounding ability acquired through Auto-AEG training also improves performance on the conventional SED task it extends can further solidify our claims. We test this on **DESED** [25], a standard domestic- environment SED benchmark with a fixed 10-class vocabulary, evaluated under event-level F1 and precision at IoU=0.5 together with mean IoU over matched events. All models receive the same List-All prompt; no SED-specific head or training signal is used, so any change reflects the grounding policy itself.

Model	mIoU	ev_F1	ev_P
Qwen3-Omni-30B	0.509	0.254	0.508
+ SFT	0.428	0.245	0.409
+ SFT + GRPO	**0.606**	**0.287**	**0.607**
Qwen2.5-Omni-7B	**0.500**	0.228	0.298
+ SFT	0.343	0.189	0.298
+ SFT + GRPO	0.452	**0.263**	**0.463**

Table 7: Results on DESED. ev_F1/ev_P: event-level F1 and precision at IoU=0.5. Bold marks the best value per model family and column.

DESED confirms that Auto-AEG training improves performance on the standard SED task that grounding extends. On the primary event-level SED metrics, GRPO beats the zero-shot baseline for *both* models: event-F1 rises (Q3-Omni 0.254 → 0.287; Q2.5-Omni 0.228 → 0.263) and event-Precision rises more sharply (Q3-Omni 0.508 → 0.607; Q2.5-Omni 0.298 → 0.463). The gain is precision-driven—GRPO commits to fewer, better-localized events rather than denser predictions—which is exactly the behavior the interval-aware reward is designed to elicit. For the larger Q3-Omni the improvement extends to overall matched-IoU (0.509 → 0.606); for the smaller Q2.5-Omni matched-IoU stays near the zero-shot level, consistent with the cross-generation finding that the 30B model exploits the reward signal more fully. Because DESED clips are only 10 s—well within the 30 s encoder window—these gains reflect boundary-localization improvement rather than any long-context advantage.

The stage-level breakdown shows that GRPO, not SFT, drives the gains. SFT alone regresses on the event metrics for both models (Q3-Omni event-F1 0.254 → 0.245, precision 0.508 → 0.409; Q2.5-Omni event-F1 0.228 → 0.189): the synthetic cold-start teaches the open-vocabulary List-All format, but its synthetic acoustics do not match DESED's real domestic recordings. GRPO then recovers and surpasses zero-shot on event-F1 and event-Precision for both models. This recovery—turning imperfect real-audio pseudo-labels into a usable reward signal—is precisely the role GRPO plays in the two-stage design, and indicates that audio event grounding and sound event detection rest on a shared temporal-localization capability that the pipeline strengthens: the grounding policy trained on open-vocabulary data transfers back to the closed-set SED task it extends.

8. Conclusion

We presented Auto-AEG, a scalable pipeline for automatically constructing open-vocabulary onset/offset supervision, paired with AEGBench, an independently annotated, difficulty-stratified benchmark for evaluating LALM temporal grounding. Fine-tuning on Auto-AEG data with a two-stage SFT cold-start and interval-aware GRPO framework yields +73.9% relative mIoU for Qwen3-Omni-30B and +23.1% for Qwen2.5-Omni-7B over zero-shot baselines; GRPO consistently amplifies the larger model across all metrics, while on the smaller 7B model it trades recall for sharper boundary precision. AEGBench's independence from the training pipeline—enforced through energy-contrast filtering, multi-phase label cleaning, and human correction—ensures that observed gains reflect genuine temporal grounding rather than annotation-pipeline imitation. The Auto-AEG

pipeline can in principle be applied to any large public audio collection, establishing a viable data-side foundation for continued advances in LALM temporal capability.

9. Limitations

The Auto-AEG Stage 1 data is synthesized from FreeSound clips, which introduce a domain mismatch relative to naturally-recorded real-world audio. While GRPO partially corrects for this by exposing the model to real audio, the Stage 1 SFT checkpoint is initialized on synthetic distributions that differ from real recordings in room acoustics, source interaction, and background statistics.

Stage 2 pseudo-labels contain localization errors from PE A-Frame (40 ms frame resolution ($\pm$20 ms boundary uncertainty)) and potential label errors from Gemini (hallucinated or imprecise descriptions). GRPO is noise-tolerant to these errors, but a more comprehensive noise analysis—e.g., measuring pseudo-label quality against held-out human timestamps—would provide stronger guarantees.

Audio recordings with long duration remain a challenge for most LALMs, which employ Whisper-based audio encoders that are limited to 30 s per encoding window.

Acknowledgments

References

[1] Orevaoghene Ahia, Martijn Bartelds, et al. "BLAB: Brutally Long Audio Bench". In: *arXiv preprint arXiv:2505.03054* (2025).

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. "Localizing moments in video with natural language". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5803–5812.

[3] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. "BEATs: Audio Pre-Training with Acoustic Tokenizers". In: *International Conference on Machine Learning (ICML)*. 2023.

[4] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. *Qwen2-Audio Technical Report*. 2024. arXiv: 2407.10759 [eess.AS]. URL: https://arxiv.org/abs/2407.10759.

[5] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. "Qwen2-Audio Technical Report". In: *arXiv preprint arXiv:2407.10759* (2024).

[6] 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". In: *arXiv preprint arXiv:2311.07919* (2023).

[7] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. "Pengi: An Audio Language Model for Audio Tasks". In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023, pp. 18090–18108.

[8] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. "FSD50K: An Open Dataset of Human-Labeled Sound Events". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 829–852.

[9] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. "Tall: Temporal activity localization via language query". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5267–5275.

[10] Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang-gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. *Audio Flamingo Next: Next-Generation Open Audio-Language Models for Speech, Sound, and Music*. 2026. arXiv: 2604.10905 [cs.SD]. URL: https://arxiv.org/abs/2604.10905.

[11] Yuan Gong, Yu-An Chung, and James Glass. "Ast: Audio spectrogram transformer". In: *arXiv preprint arXiv:2104.01778* (2021).

[12] Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. "Listen, Think, and Understand". In: *International Conference on Learning Representations*. 2024.

[13] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. "Acoustic scene classification in DCASE 2020 Challenge: generalization across devices and low complexity solutions". In: *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*. 2020, pp. 56–60. URL: https://arxiv.org/abs/2005.14623.

[14] Ailin Huang et al. *Step-Audio: Unified Understanding and Generation in Intelligent Speech Interaction*. 2025. arXiv: 2502.11946 [cs.CL]. URL: https://arxiv.org/abs/2502.11946.

[15] Kimi Team. *Kimi-Audio Technical Report*. https://github.com/MoonshotAI/Kimi-Audio. Open-source 7B audio language model with native temporal grounding support. 2025.

[16] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition". In: *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 28 (Nov. 2020), pp. 2880–2894. ISSN: 2329-9290. DOI: 10.1109/TASLP.2020.3030497. URL: https://doi.org/10.1109/TASLP.2020.3030497.

[17] Jie Lei, Tamara L Berg, and Mohit Bansal. "Detecting moments and highlights in videos via natural language queries". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 11846–11858.

[18] Zijia Liu, Peixuan Han, Haofei Yu, Haoru Li, and Jiaxuan You. *Time-R1: Towards Comprehensive Temporal Reasoning in LLMs*. 2025. arXiv: 2505.13508 [cs.CL]. URL: https://arxiv.org/abs/2505.13508.

[19] 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". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), pp. 3339–3354.

[20] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. "Query-dependent video representation for moment retrieval and highlight detection". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 23023–23033.

[21] Hokuto Munakata, Taichi Nishimura, Shota Nakada, and Tatsuya Komatsu. "Language-based Audio Moment Retrieval". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2025.

[22] Shota Nakada, Taichi Nishimura, Hokuto Munakata, and Tatsuya Komatsu. "CASTELLA: Long Audio Dataset with Captions and Temporal Boundaries". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2026.

[23] Paul Primus, Florian Schmid, and Gerhard Widmer. *TACOS: Temporally-aligned Audio CaptiOnS for Language-Audio Pretraining*. 2025. arXiv: 2505.07609 [cs.SD]. URL: https://arxiv.org/abs/2505.07609.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning* (2021), pp. 8748–8763.

[25] Romain Serizel, Nicolas Turpault, Hamid Eghbalzadeh, and Ankit Shah. "Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments". In: *ArXiv* abs/1807.10501 (2018). URL: https://api.semanticscholar.org/CorpusID:51864948.

[26] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. "Deepseekmath: Pushing the limits of mathematical reasoning in open language models". In: *arXiv preprint arXiv:2402.03300* (2024).

[27] Luoyi Sun, Xiao Zhou, Zeqian Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. *SpotSound: Enhancing Large Audio-Language Models with Fine-Grained Temporal Grounding*. 2026. arXiv: 2604.13023 [cs.SD]. URL: https://arxiv.org/abs/2604.13023.

[28] 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: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=14rn7HpKVk.

[29] Qwen Team. "Qwen2.5-Omni Technical Report". In: *arXiv preprint arXiv:2503.20215*. 2025.

[30] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis". In: *Workshop on Detection and Classification of Acoustic Scenes and Events*. 2019.

[31] Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matt Le, Piotr Dollár, Christoph Feichtenhofer, Ann Lee, and Wei-Ning Hsu. *Pushing the Frontier of Audiovisual Perception with Large-Scale Multimodal Correspondence Learning*. 2025. arXiv: 2512.19687 [cs.SD]. URL: https://arxiv.org/abs/2512.19687.

[32] Hualei Wang, Yiming Li, Shuo Ma, Hong Liu, and Xiangdong Wang. "Listening Between the Frames: Bridging Temporal Gaps in Large Audio-Language Models". In: *arXiv preprint arXiv:2511.11039* (2025).

[33] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. *Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation*. 2024. arXiv: 2211.06687 [cs.SD]. URL: https://arxiv.org/abs/2211.06687.

[34] Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. *AudioTime: A Temporally-aligned Audio-text Benchmark Dataset*. 2024. arXiv: 2407.02857 [eess.AS]. URL: https://arxiv.org/abs/2407.02857.

[35] Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. *PicoAudio: Enabling Precise Timestamp and Frequency Controllability of Audio Events in Text-to-audio Generation*. 2024. arXiv: 2407.02869 [eess.AS]. URL: https://arxiv.org/abs/2407.02869.

[36] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. *Qwen3-Omni Technical Report*. 2025. arXiv: 2509.17765 [cs.CL]. URL: https://arxiv.org/abs/2509.17765.

[37] Xuenan Xu, Heinrich Dinkel, Mengyue Wu, and Kai Yu. "Text-to-Audio Grounding: Building Correspondence Between Captions and Sound Events". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, pp. 606–610.

[38] Xuenan Xu, Ziyang Ma, Mengyue Wu, and Kai Yu. "Towards Weakly Supervised Text-to-Audio Grounding". In: *IEEE Transactions on Multimedia* (2024).

[39] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. "Semantic conditioned dynamic modulation for temporal sentence grounding in videos". In: *Advances in Neural Information Processing Systems* 32 (2019).

[40] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. "Span-based localizing network for natural language video localization". In: *Proceedings of the 58th annual meeting of the association for computational linguistics*. 2020, pp. 6543–6554.

[41] Zhen Zuo, Bing Shuai, Gang Wang, Xiao Liu, Xingxing Wang, Bing Wang, and Yushi Chen. "Convolutional recurrent neural networks: Learning spatial dependencies for image representation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2015, pp. 18–26.

A. Training Prompt Template

System prompt (shared by SFT and GRPO).

```
You are an audio temporal grounding assistant. Listen carefully to the audio
and answer questions about when specific sound events occur. Think step by step
inside <think> tags, then give your final answer inside <answer> tags.
```

User query (List-All).

```
The audio contains the sound event: "{category}". List ALL time intervals
(start, end in seconds) when this event occurs. Reply ONLY with a JSON array
inside <answer> tags, e.g. <answer>[[0.5, 2.1], [5.0, 7.3]]</answer> or
<answer>[]</answer> if not present.
```

Gold response with reasoning (SFT target).

```
<think>
I need to find all time intervals when {category} occurs in this audio
(total duration: {duration}s). Listening carefully, I can identify
{n} occurrences: {parts}.
</think>
<answer>{ans}</answer>
```

Figure 2: Prompt template shared by SFT and GRPO. The system prompt elicits a `<think>` reasoning trace before the `<answer>`. During SFT the model is trained on gold responses whose chain-of-thought enumerates every occurrence (the single-occurrence target is identical with "one occurrence"); during GRPO the same system prompt and query are used, but the response is sampled and scored by the interval-aware reward rather than supervised. A no-rejection query variant (omitting the `<answer>[]` option) is used for clips where the target event is guaranteed present.

B. Stage 1 Synthesis Details

Each synthetic clip is constructed as follows:

(1) randomly select a target label from the pool;

(2) sample 1–5 non-overlapping segments of that label (inter-segment gap uniformly drawn from $[0.5, 4.0]$ s);

(3) optionally insert 0–2 distractor events of a *different* label, which may overlap freely;

(4) mix onto a Gaussian background with foreground SNR uniformly drawn from $[10, 20]$ dB;

(5) set total clip duration to $[10, 30]$ s at 16 kHz.

The distribution of target occurrence counts is skewed toward multiple occurrences: 1 event (20%), 2 (30%), 3 (25%), 4 (15%), 5 (10%). Clips are stored at 16 kHz in FLAC format.

Each clip is paired with a List-All query:

> *"The audio contains the sound event: `[label]`. List ALL time intervals when this event occurs. Format your answer as a JSON array of [start, end] pairs in seconds."*

with response `<answer>[[s1,e1],[s2,e2],...]</answer>`. A short `<think>` prefix is prepended during training to encourage step-by-step temporal reasoning.

C. Stage 2 Annotation Details

The Stage 2 pipeline (also used for AEGBench Phase 1) runs three steps per clip.

Label identification. The audio is chunked into 10 s segments—chunking mitigates the tendency of LALMs to miss events in long recordings—and each chunk is sent to Gemini, which returns the clearly audible events in JSON (Figure 3). The prompt requests at most eight semantically orthogonal labels (1–3 words, lowercase), prefers the most specific descriptor over a parent category (*dog barking* over *animal sound*), and abstains from generic ambient labels (*background noise*, *music*, *wind*) unless such a sound is itself the sole dominant event. Chunk-level labels are pooled and lightly canonicalized (character normalization, deduplication, stop-list removal) into a per-clip inventory.

Event-type classification. Each label is classified as *continuous* (sustained events such as engine noise or rain) or *discrete* (transient events such as a knock or a bark) via a text-only Gemini query (see Figure 4), cached per unique label string.

Temporal localization. PE A-Frame computes per-frame audio-text similarity at ≈40 ms resolution over the full clip; frames whose score exceeds 0.5 are marked active and consecutive active frames are merged into onset/offset spans. For continuous labels, adjacent spans separated by at most 0.5 s are further merged to suppress brief dropouts; for discrete labels no merging is applied, so each peak yields a distinct occurrence. Labels for which no active frame is found are discarded, ensuring every retained span reflects a confident detection.

Global label cleaning. Because labels are coined independently per clip, near-synonyms (*car engine*, *engine noise*, *vehicle engine*) may coexist. We collect all unique labels, embed each with CLAP [33], and cluster the embeddings by k-means with $k = \lceil N/10 \rceil$ (on average $\leq$10 labels per cluster). Each cluster, with per-label occurrence counts, is sent to Gemini, which assigns every label *keep* (specific and non-redundant), *drop* (generic or duplicating a more specific label in the same cluster), or *rename* (replace by a provided canonical form; Figure 4, bottom). The resulting global mapping is applied to every record.

```
Identify clearly recognizable sound events in this audio clip.
Reply ONLY with JSON: {"sounds": ["label1", "label2", ...]}.
Return at most 8 labels.
Important rules:
(1) Labels should be semantically orthogonal: do not output multiple labels that
    point to the same event family. For example, choose only one of
    "bird chirping", "bird singing", "bird vocalization"; choose only one of
    "bus", "motor vehicle", "vehicle".
(2) Prefer the most specific audible event label, not its parent category.
(3) Do NOT output generic ambience-only labels such as background noise, noise,
    human voice, sound effect, music, water, or wind, unless that generic sound
    is itself the single dominant event.
(4) If unsure whether a label is distinct or dominant enough, abstain. Prefer
    fewer labels.
(5) Use short common English phrases, 1-3 words, lowercase style.
Reply with compact minified JSON only. Do not use markdown, code fences,
explanations, or extra keys.
```

Figure 3: Label-identification prompt, sent once per 10 s audio chunk; labels from all chunks are unioned into the per-clip inventory.

Event-type classification (per unique label, text-only).

```
Is the sound "{label}" continuous (sustained over time, e.g. engine noise, rain,
music, wind) or discrete (short events, e.g. knock, clap, beep, gunshot)?
Reply with only one word: continuous or discrete.
```

Cluster-level label cleaning (per k-means cluster). Here {`items`} is a comma-separated list of `"label" (n=count)` pairs.

```
You are an audio taxonomy expert. Here is a cluster of sound labels: [{items}].
For each label decide: keep (it is specific and useful), drop (it is
generic/redundant), or rename (provide a better canonical form).
Reply ONLY with JSON: {"decisions": {
    "label": {"action": "keep"|"drop"|"rename", "rename_to": "new_name_or_null"}
}}.
Lowercase, 1-3 word phrases.
```

Figure 4: Event-type classification (top) and global label-cleaning (bottom) prompts. The classification result controls whether adjacent PE A-Frame spans are merged (continuous) or kept separate (discrete).

D. AEGBench Construction Details

Source distribution. After filtering, the four sources retain 2,230 (AudioSet Strong Labels), 954 (FSD50K eval), 234 (BBC Sound Effects, with single-source clips excluded to increase scene diversity), and 9 (YouTube Life Sounds) items.

Quality filters. Each candidate must satisfy: **energy contrast** $\geq 12\,\mathrm{dB}$ (the target event is measurably louder than the background); **active ratio** $\in [0.10, 0.85]$ (neither predominantly silent nor a continuous event with no reference background); **active duration** $\in [0.5, 60.0]\,\mathrm{s}$; and a **per-category cap** of at most 40 items per fine-grained class.

Hard-case tagging. Difficulty tags (not mutually exclusive) are assigned by the following rules.

1. **Polyphonic Overlap**: 2 or more clips of different categories overlap by more than 20% of the shorter clip's duration.
2. **Repeated Occurrence**: a category has 3 or more detected spans.
3. **Long Duration**: a span exceeds 30 s.
4. **Semantic Ambiguity**: the CLAP cosine similarity between two distinct labels lies in $[0.40, 0.85]$.
5. **Low-Contrast**: the energy-contrast score lies in $[12.0, 28.0]\,\mathrm{dB}$—above the 12 dB salience threshold but at its low end, so the event is only moderately louder than its background and its boundary is harder to localize precisely.
6. **Gradual Onset/Offset**: the pre-roll or post-roll of the target segment is longer than 500 ms, or the label contains a perceptual-gradation keyword (*approaching*, *fading*, *building up*, *passing by*).

For stratified diagnostics, items within each category are ranked by energy contrast and the top 100 retained.