`iKnow-audio`: Integrating Knowledge Graphs with Audio-Language Models

Anonymous ACL submission

Abstract

Contrastive audio-language models are learned by semantically aligning different modalities in a shared embedding space. Existing research shows that zero-shot classification performance is sensitive to language nuances and prompt formulation. In addition, learned artifacts and spurious correlations from noisy pretraining often lead to semantic ambiguity in label interpretation. While recent work has explored few-shot prefix tuning methods, adapters, and prompt engineering strategies to mitigate these issues, the use of structured prior knowledge remains largely unexplored. In this work, we enhance CLAP predictions using structured reasoning over a knowledge graph (KG). We construct a large, audio-centric KG that encodes ontological relations comprising semantical, causal, and taxonomic connections reflective of everyday sound scenes and events. A systematic analysis of retrieval performance across major publicly available audio collections demonstrates that symbolic knowledge enables robust semantic grounding for contrastive audio-language models. This improvement is further supported by embedding visualizations of CLAP before and after incorporating the KG.

1 Introduction

In recent years, self-supervised and multimodal models such as contrastive language-audio pretraining (CLAP) (Elizalde et al., 2023) have shown impressive performance in audio understanding tasks by leveraging large-scale contrastive learning between audio and natural language descriptions. While excelling at capturing general semantic correspondences, these models often lack a deeper understanding of the relational and contextual structure of real-world sound events. Common deficiencies include disambiguating acoustically similar sounds, modeling co-occurrence patterns or hierarchical relationships, and a lack of commonsense grounding necessary for reasoning about sounds in novel contexts. Additionally, the performance of these models relies heavily on prompt engineering. Indeed, previous work has shown that changes in prompt wording and formatting can substantially affect performance in zero-shot audio classification tasks (Olvera et al., 2024).

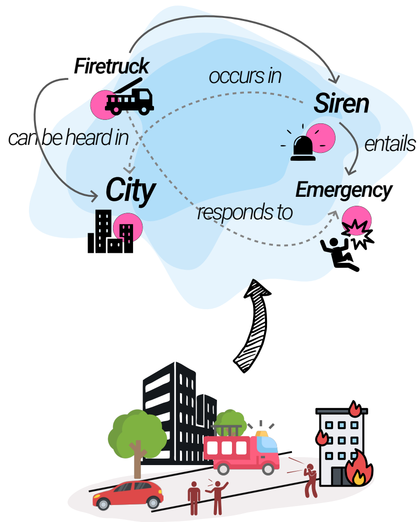

Figure 1: Audio understanding requires contextual and background knowledge, which can be represented by audio knowledge graphs (AKG).

Understanding real-world sounds often requires contextual and background knowledge. For example, in the scenario illustrated in Figure 1, recognizing that the sound of *sirens* may indicate emergency vehicles, which are often associated with *accidents*, *fires*, or *emergencies*, and that sirens frequently co-occur with *engine noise*, *people shouting*, or *braking sounds*. Such relationships go beyond mere labels; they reflect structured, situational knowledge that is paramount for accurate interpretation. While datasets exist for sound events, they lack a structured semantic representation of such interconnections. We address this gap by constructing a general-purpose Audio Knowledge Graph (AKG) that captures rich, multi-relational information about sound-producing entities, their categories, and co-occurrence patterns. This structured knowledge is vital for enabling downstream models to reason about ambiguous or acoustically similar sounds, particularly in low-resource or zero-shot settings where such priors are otherwise absent.

While a knowledge graph is a powerful source of relational knowledge, querying it directly using symbolic methods (e.g., rule-based lookup or SPARQL-style queries) is limited to exact matches and fails to generalize or infer new knowledge beyond what's explicitly encoded. Knowledge embedding models (KEMs) address this limitation by mapping entities and relations into continuous vector spaces, allowing for: generalization to unseen or sparse triples through latent similarity, robust reasoning under uncertainty or label noise, efficient link prediction (e.g., inferring *yelping* as a plausible child category of *dog* even if not explicitly stated).

To leverage the capability of CLAP while addressing its limitations, we propose to refine CLAP predictions with the KEMs obtained from the AKG. Importantly, for the text input, we encode only the class labels, instead of extended prompts, such as "*This is a sound of {}*" or "*A recording of {}*". This minimizes the efforts on prompt engineering and allows us to focus on improving CLAP predictions with factual knowledge about sounds.

In summary, we present the following contributions: (1) **AKG**: A comprehensive audio knowledge graph that encodes rich relational semantics among everyday sounds. (2) **CLAP-KG**: A novel pipeline for refining CLAP predictions using a knowledge embedding model trained on AKG. (3) Systematic zero-shot evaluation on six benchmark datasets, showing consistent improvements over baseline CLAP.

2 Related Work

Multimodal and Domain-Specific Knowledge Graphs Conventional knowledge graphs are typically limited to the textual space, restricting their efficacy on other modalities (Hogan et al., 2021). Recent research has aimed to overcome this limitation by integrating cross-modal knowledge. Wang et al. (Wang et al., 2023) first constructed a multimodal KGs incorporating text, image, video, and audio modalities, supported by extensively annotated datasets. A unified pipeline was proposed in (Gong et al., 2024) to help construct multimodal KGs. Wei et al. built domain-specific KGs by connecting medical images and their related biomedical concepts (Wei et al., 2024). To the best of our knowledge, there are currently no knowledge graphs representing rich relational semantics among everyday sounds.

Vision-Language Models with KGs Due to the inherent hallucination artifacts of large language models (LLMs), there is a trend to use factual knowledge to enhance reasoning with vision-language models. Liu et al. (Liu et al., 2025) proposed a method that enhances LLMs' multimodal reasoning abilities through an integrated KG constructed via vision-language alignment with cross-modal similarity recalibration. Similarly, Li et al. (Li et al., 2023) proposed GraphAdapter, a fine-tuning framework leveraging dual KGs to improve vision-language understanding. A cross-modal alignment module was introduced in (Lee et al., 2024) to align knowledge from images and text in vision-language fine-tuning. Gao et al. (Gao et al., 2025) introduced a retrieve-and-rerank framework for KG-augmented contrastive Language-Image Pre-Training (CLIP).

Leveraging KGs for Audio Despite the advances in vision-language KGs, the audio modality remains relatively under-explored. Penamakuri et al. (Penamakuri et al., 2025) introduced Audiopedia, a framework for audio-based question answering augmented with external knowledge. However, the authors' approach continues to rely on text-based KGs, either by enhancing prompts or through intermediate automatic speech recognition, rather than constructing an audio-specific KG. The approach proposed in this paper is closely related to (Gao et al., 2025). Due to the paradigm of CLIP, the full pipeline in that work still depends heavily on prompt engineering. Unlike prior methods, our approach merely relies on the class labels as the text prompt, allowing us to focus on the semantic connection between modalities and use the the AKG to enhance reasoning.

3 Audio-language Models with KG Reasoning

In this section, we detail the construction of our audio-centric knowledge graph, the training of knowledge graph embedding models, and their integration into a reasoning pipeline designed to refine the zero-shot predictions of the CLAP model. Our methodology, as shown in Figure 2, is adaptable to

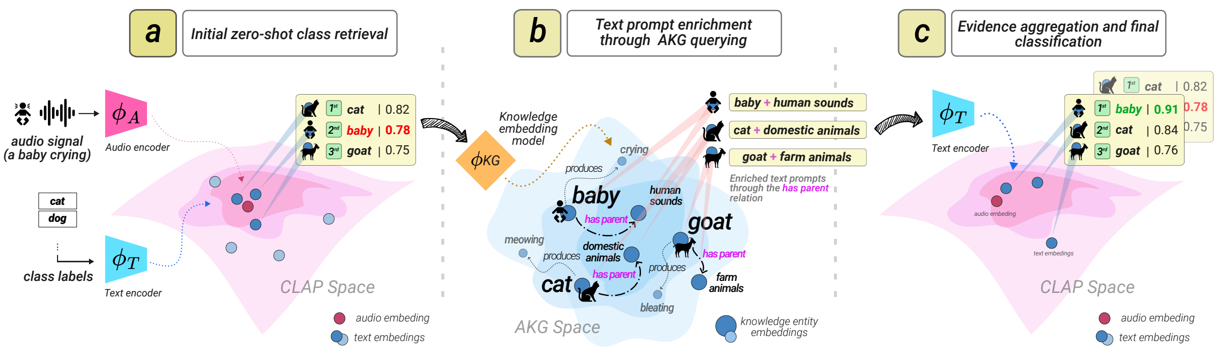

Figure 2: Our pipeline enhances zero-shot audio classification via KG reasoning. (a) CLAP initially misranks the correct label (e.g., *baby*) due to acoustic ambiguity with other labels. (b) We query an audio-centric KG using top-k predictions to retrieve related concepts via relevant relations (e.g., `has parent`). (c) Enriched prompts are compared with the audio embedding, and similarity scores are aggregated to re-rank predictions, this time correctly identifying *baby* as the top label. This refinement demonstrates the utility of structured symbolic knowledge for disambiguating acoustic scenes and improving interpretability.

any audio-language model featuring aligned audio and text encoders.

3.1 Knowledge Embedding Model

To enable structured reasoning over audio-centric relationships, we employ knowledge graph embedding models that learn vector representations for entities and relations. These embeddings support link prediction, allowing the model to infer plausible but unobserved relations between audio concepts.

We represent the knowledge graph as $\mathcal{G} = (\mathcal{E}, \mathcal{R})$, where $\mathcal{E}$ denotes the set of entities (e.g., *siren*, *barking*) and $\mathcal{R}$ the set of relation types (e.g., `belongs to class`, `co-occurs with`). Each factual statement is encoded as a triple $(h, r, t) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, where h is the head entity, r the relation, and t the tail entity. For example, the triple (*dish clinking*, `occurs in`, *kitchen*) captures a spatial context in which the sound typically appears.

We define a scoring function $\phi_{\text{KG}} : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \to \mathbb{R}$, which assigns a plausibility score to a given triple (h, r, t). In our zero-shot classification pipeline, this function is primarily used for link prediction, specifically tail prediction, where, given a head entity h and relation r, we rank candidate tail entities $t \in \mathcal{E}$ based on their plausibility. Higher scores indicate greater semantic compatibility, enabling the discovery of relevant or missing connections between audio concepts.

To model these interactions, we experiment with several knowledge embedding approaches implemented via the PyKEEN library. These include: (1) **TransE** (Bordes et al., 2013), which models relations as translations in the embedding space. (2) **TransH** (Wang et al., 2014) and **TransR** (Lin et al., 2017), which extend TransE by introducing relation-specific projection spaces ; (3) **ComplEx** (Trouillon et al., 2016), which leverages complex-valued embeddings to model asymmetric relations; (4) **RotatE** (Sun et al., 2019), which represents each relation as a rotation in the complex vector space $\mathbb{C}^d$; and (5) **GCN** -based (graph convolutional network) models (Schlichtkrull et al., 2017), which propagate information through the graph structure via message passing.

In this work, we adopt the RotatE model due to its strong empirical performance on the AKG dataset (see Section 5.5). RotatE embeds entities and relations in a complex vector space $\mathbb{C}^d$, and each relation is modeled as a rotation in that space. The score of a triple (h, r, t) is given by:

$$\phi_{\text{KG}}(h, r, t) = - \|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\|_2 , \qquad (1)$$

where $\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{C}^d$ are the embeddings of the head, relation, and tail, respectively, and $\circ$ denotes the element-wise (Hadamard) product. A higher score indicates a more plausible triple.

This scoring mechanism enables structured reasoning over multi-relational knowledge, which we exploit to retrieve semantically related entities and refine CLAP's initial predictions via link prediction.

3.2 Zero-Shot Classification with CLAP

We leverage CLAP (Elizalde et al., 2023), a pre-trained model that embeds audio and text into a shared representation space. This enables zero-shot audio classification by computing similarity scores between audio inputs and candidate label embeddings.

Let $\mathcal{A}$ denote the space of input audio signals and $\mathcal{L}$ the space of textual labels. Given a set of target class labels $C = \{c_1, \ldots, c_N\} \subset \mathcal{L}$ and an input audio sample $a \in \mathcal{A}$, CLAP maps both modalities into a joint embedding space via an audio encoder $\phi_{\mathrm{A}} : \mathcal{A} \to \mathbb{R}^d$, and a text encoder $\phi_{\mathrm{T}} : \mathcal{L} \to \mathbb{R}^d$.

CLAP formulates classification as a nearest-neighbor retrieval task (Figure 2 (a)), where the predicted label $\hat{c} \in C$ is obtained by maximizing cosine similarity:

$$\hat{c} = \arg\max_{c \in C} \mathrm{sim}\left(\phi_{\mathrm{A}}(a), \phi_{\mathrm{T}}(c)\right), \qquad (2)$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity. We denote the top-k retrieved labels as:

$$C_k = \{c^{(1)}, \ldots, c^{(k)}\}, \quad \text{ranked by similarity.}$$

3.3 Enhancing CLAP Inference with AKG

To enhance interpretability and robustness, we refine the predictions C_k via symbolic reasoning over $\mathcal{G}$. This produces enriched, context-aware prompts that reflect the semantic neighborhood of each class. This process is depicted in Figure 2 (b).

Link Prediction To enrich top-k CLAP predictions with structured knowledge, we perform link prediction using the trained knowledge embedding model ϕ_{KG}. Given a predicted class label $c \in C_k$, we use ϕ_{KG} to infer the most semantically plausible tail entities $t \in \mathcal{E}$ connected to c via a curated subset of informative relations $\mathcal{R}_q \subset \mathcal{R}$. These predicted tails serve as contextual signals to refine and expand the textual prompts used for similarity computation within the CLAP model.

Contextual Prompt Expansion For each top prediction $c \in C_k$, we query the knowledge graph to retrieve candidate tail entities connected via informative relations:

$$\mathcal{T}_c = \{(c, r, t) \in \mathcal{T} \mid r \in \mathcal{R}_q\},$$

where $\mathcal{R}_q \subset \mathcal{R}$ is a curated set of relations used for semantic enrichment (e.g., `produces`).

Using the knowledge embedding model ϕ_{KG}, we rank tail candidates $t \in \mathcal{E}$ for each relation $r \in \mathcal{R}_q$ based on their plausibility in completing the triple (c, r, t). We select the top-m most plausible tails:

$$\mathcal{T}_c^{\mathrm{top}} = \{t_1^*, \ldots, t_m^*\},$$

where $t_i^* \in \arg\max_{t \in \mathcal{E}} \mathrm{score}(c, r, t; \phi_{\mathrm{KG}})$, and $\mathrm{score}(\cdot)$ denotes the plausibility score assigned by ϕ_{KG}.

To generate enriched prompts, we concatenate each class label c with its associated tail entities t_i^*. For example, prompts can take the form:

$$p_{c,t_i^*} = \mathtt{concat}(c, t_i^*).$$

Let $P_c = \{p_{c,t_1^*}, \ldots, p_{c,t_m^*}\}$ be the set of knowledge-enriched prompts associated with class c.

Scoring with Enriched Prompts Each enriched prompt $p \in P_c$ is encoded using the CLAP text encoder ϕ_{T}, and scored against the input audio $a \in \mathcal{A}$ via cosine similarity:

$$s(p) = \mathrm{sim}\left(\phi_{\mathrm{A}}(a), \phi_{\mathrm{T}}(p)\right). \qquad (3)$$

This yields a refined similarity score for each knowledge-augmented prompt, enabling re-ranking of the initial predictions C_k based on semantically enriched textual context.

Aggregation and Re-ranking To consolidate evidence from both the original label and its knowledge-augmented prompts, we aggregate their similarity scores into a single score per class (Figure 2 (c)).

For each class $c \in C_k$, let $s(c) = \mathrm{sim}(\phi_{\mathrm{A}}(a), \phi_{\mathrm{T}}(c))$ denote the original CLAP score, and $\{s(p) \mid p \in P_c\}$ the scores of its enriched prompts. We define the aggregated score $\tilde{s}(c)$ using a log-sum-exp fusion:

$$\tilde{s}(c) = \log\left(\exp(s(c)) + \sum_{p \in P_c} \exp(s(p))\right). \qquad (4)$$

This operation softly pools evidence across the original and contextualized prompts. The final class prediction is given by:

$$\tilde{c} = \arg\max_{c \in C_k} \tilde{s}(c). \qquad (5)$$

A detailed description of the algorithm is provided in Appendix 1.

4 Knowledge Graph Construction

Sound events are ubiquitous and seldom occur in isolation. They are situated within broader contexts that encompass temporal dynamics, causal relations, environmental cues, perceptual attributes, and even human intent. Capturing such relationships is essential for integrating commonsense knowledge, easing robust inference and better generalization in audio tasks. To move beyond conventional classification paradigms, we construct a domain-specific Audio-centric Knowledge Graph that encodes these relational semantics among everyday sounds.

Unlike general-purpose KGs such as DBpedia, ConceptNet, and Wikidata, which offer limited coverage of everyday sounds and lack fine-grained audio semantics and perceptual grounding, our AKG is tailored for auditory scenes, enabling symbolic reasoning aligned with audio-language models.

We construct an Audio knowledge graph (AKG) to encode structured knowledge about sound events and their semantic and contextual properties. We derive this graph from standardized sound event labels aggregated across over 27 publicly available datasets, as cataloged in the SALT taxonomy (Stamatiadis et al., 2024). Our AKG includes entities such as sound-producing sources (e.g., *dog*, *engine*), sound events (e.g., *barking*, *idling*), and higher-level categorical labels (e.g., *domestic animal*, *vehicle*).

The schema comprises 8 high-level relation categories, each reflecting distinct aspects of auditory context. These categories guide the generation of plausible triples (head, relation, tail), where the head is a standardized sound event label and the relation contextualizes its link to the tail concept. We create a triple dataset with relations like `has parent` and `occurs in` for training knowledge graph embeddings (see Section 3.1). The full relation schema is detailed in Appendix A.1.

We construct the knowledge graph by generating triples from two sources: (1) the hierarchical structure of the SALT taxonomy, and (2) large language model (LLM)-generated triples based on SALT labels. Using Mistral-7B-Instruct, we produce an initial set of 51,254 triples, which undergo a two-stage filtering process—LLM-based plausibility checks followed by manual refinement. This yields a curated set of 20,387 unique, high-quality triples. Prompt templates used for triple generation are detailed in Appendix A.5, while summary statistics are provided in Appendix A.2.

5 Evaluation

5.1 Datasets

We evaluate our approach on six benchmark datasets designed for single-class or multi-label environmental sound classification:

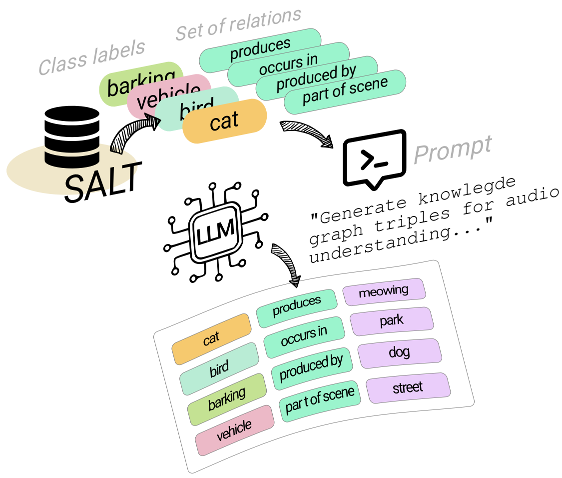

Figure 3: Generation of knowledge triples from LLMs.

ESC50 (Piczak): A dataset of 2,000 labeled 5-second audio clips spanning 50 environmental sound classes. **UrbanSound8K** (Salamon et al., 2014): Comprises 8,732 labeled audio excerpts, each with a duration of up to 4 seconds, across 10 urban sound categories. **TUT2017** (Mesaros et al., 2016): Contains 6,300 10-second recordings representing 15 distinct acoustic scenes. **FSD50K** (Fonseca et al., 2022): A collection of 51,197 variable-length audio clips (0.3–30 seconds) from Freesound, annotated across 200 classes. **AudioSet** (Gemmeke et al., 2017): A large-scale dataset with over 2 million 10-second YouTube clips, covering 527 diverse sound categories. **DCASE17-T4** (Mesaros et al., 2017): A curated subset of AudioSet focusing on 17 warning and vehicle sound classes, consisting of 52,763 10-second clips. We utilize all cross-validation folds for ESC50, US8K, and TUT2017, and test sets for AudioSet (20,371), FSD50K (20,462), and DCASE17 (488).

5.2 Prompt Format

We use only the raw class labels from SALT, formatted in lowercase with underscores replaced by spaces (e.g., *dog_barking* $\rightarrow$ *dog barking*). This deliberate choice avoids the variability and required dataset-specific tuning typically introduced by prompt engineering. This setup allows to isolate the contribution of structured knowledge in refining CLAP's predictions, withut confounding effects from prompt engineering. Although not optimized for best-case accuracy, it offers a clean and consistent basis for evaluating the impact of knowledge-based reasoning for zero-shot audio classification.

5.3 Metrics

We use two metrics to measure the performance across datasets.

Hit@K: For a given query, hit@k computes the ratio of ground truth that has been retrieved among the top K candidates.

Mean reciprocal rank (MRR): The average of the reciprocal ranks of ground truth across multiple queries. For each query, the reciprocal rank is the inverse of the position at which the ground truth appears in the ranked list.

AKG Model Training To learn structured representations over our audio-centric knowledge graph, we trained a suite of knowledge embedding models using the `PyKEEN` library (Ali et al., 2021). We evaluated six established models: TransE (Bordes et al., 2013), TransH (Wang et al., 2014), TransR (Lin et al., 2017), ComplEx (Trouillon et al., 2016), R-GCN (Schlichtkrull et al., 2017), and RotatE (Sun et al., 2019). For each model, we conducted a grid search over the following hyperparameters: batch size (values in $\{2^8, 2^9, 2^{10}, 2^{11}, 2^{12}\}$), learning rate (in $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$), and embedding dimensionality ($\{64, 128, 256\}$). Training was performed on two variants of the knowledge graph: (i) a *noisy* version composed of raw triples extracted from sound event labels without further refinement, and (ii) a *clean* version derived through LLM-based plausibility verification and manual post-processing to remove duplicates, spurious entries and inconsistencies in label granularity.

5.4 Results

We start by comparing the different embedding models, then look at the zero-shot audio classification results.

5.4.1 Embedding Models

Table 1 presents a comparison of models envisaged for our AKG embedding. We evaluated each model on link prediction tasks, comparing performance under both the initial noisy and cleaned versions.

Noisy vs Clean Settings Transitioning from the noisy to the clean graph yields substantial performance gains for all models, underscoring the importance of post-processing triples. Notable improvements include TransH's MRR rising from 11.8 to 26.1 and R-GCN's from 30.0 to 41.7. This supports the notion that spurious triples and inconsistencies in entity labeling can obscure latent relational patterns crucial for learning effective embeddings.

Model	Hits@1	Hits@3	Hits@5	Hits@10	MRR
Noisy graph					
TransE	1.0	<u>36.0</u>	<u>47.4</u>	<u>59.8</u>	22.2
TransH	6.0	12.1	16.7	22.5	11.8
TransR	3.4	7.1	9.6	13.3	7.1
ComplEx	<u>19.6</u>	34.3	40.9	50.5	<u>30.1</u>
R-GCN	17.4	33.8	43.9	56.7	30.0
RotatE	**37.0**	**56.9**	**64.8**	**73.2**	**49.5**
Clean graph					
TransE	1.6	40.8	50.9	60.6	24.3
TransH	17.3	28.9	35.5	43.5	26.1
TransR	7.3	15.0	18.8	25.1	13.6
ComplEx	22.7	35.1	40.1	48.2	31.3
R-GCN	<u>28.6</u>	<u>47.7</u>	<u>57.4</u>	<u>68.8</u>	<u>41.7</u>
RotatE	**46.4**	**61.9**	**67.7**	**74.0**	**56.1**

Table 1: Comparison of models on clean and noisy conditions. Retrieval results (%) in terms of hit@1, hit@3, hit@5, and MRR on the six benchmark model. Best performances are in **bold** and second-best are <u>underlined</u>.

Model-based Performance RotatE outperforms all models in both clean and noisy settings, achieving the highest MRR (56.1) and leading in all Hits@K metrics. Its performance effectively captures asymmetric and compositional relations such as `produces`, or `causes`, outperforming simpler translational models like TransE and TransH. R-GCN performs well on the clean graph due to its use of structural information but is highly sensitive to noise, where simpler models like TransE and ComplEx perform better. Despite its strengths, R-GCN slightly underperforms RotatE, possibly due to weaker handling of relation directionality or suboptimal tuning. ComplEx, effective for asymmetric relations, shows no notable gains in the clean setting, performing similarly across both conditions.

AKG Model Selection Based on this comparative analysis, we select RotatE as the backbone model for downstream knowledge reasoning/querying. Its superior link prediction capabilities ensure that the semantic augmentations introduced to CLAP are grounded in plausible, relationally informed expansions of the label space. The robustness of RotatE across both clean and noisy settings further supports its integration into our inference-time zero-shot audio classification pipeline.

5.4.2 Zero-Shot Audio Classification

Table 2 lists the best retrieval performance of CLAP and CLAP-KG on the six benchmark datasets. We observe that performance improves for all metrics over all datasets except for the hit@5 metric. This may be attributed to the semantic closeness of top-K candidates to the ground truth. Considering

Dataset	ESC50	US8K	TUT2017	FSD50K	AudioSet	DCASE17-T4
Hit@1	93.2 | **95.4**	82.5 | **85.9**	37.8 | **47.9**	61.1 | **64.0**	18.4 | **19.9**	37.7 | **45.9**
Hit@3	98.8 | 99.2	96.6 | 96.9	74.9 | **83.3**	82.8 | **84.2**	33.1 | **34.4**	77.3 | **78.5**
Hit@5	99.5 | 99.5	98.8 | 98.8	91.3 | 91.3	88.9 | 88.9	41.1 | 41.1	91.2 | 91.2
MRR	95.9 | **97.2**	89.6 | 91.5	57.7 | **65.4**	72.2 | **74.3**	26.5 | **27.7**	57.3 | **63.1**

Table 2: Retrieval results (%) in terms of hit@1, hit@3, hit@5, and MRR on the six benchmark datasets: CLAP | CLAP-KG. Performance improvement larger 1% is in **bold** and that less than 1% is underlined.

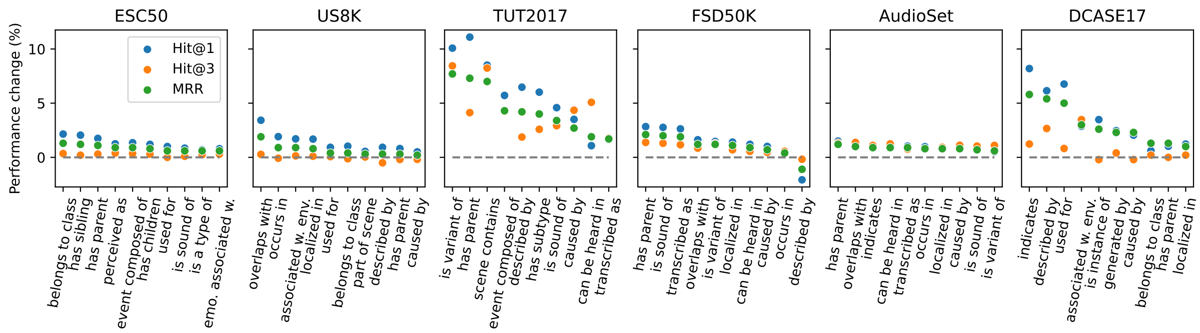

Figure 4: Performance change (%) of CLAP-KG as compared to CLAP in terms of Hit@1, Hit@3, and MRR. Only the top 10 relationships are displayed. `associated w. env.` = `associated with environment`; `emo. associated w. env.` = `emotionally associated with`.

more candidates increases the likelihood of having the ground truth, applicable for both CLAP and CLAP-KG. The most impressive improvement is the hit@1 on the TUT2017 by 11.1%, highly due to the context and background knowledge required to understand an acoustic scene. Relations like `scene contains` or `described as` disentangle the auditory scene into its sound event components.

Impact of Relations Datasets often vary in terms of context and structure, reflecting different relations among classes. To shed light on this perspective, we plot the zero-shot classification (ZSAC) performance with different relations, as shown in Figure 4. Clearly, many relations boost the performance across all datasets. `has parent` is a shared relation that works for all datasets. This is expected due to the inherent taxonomical categorization of sound events reflected in many datasets, where labels are systematically grouped into categories. The most impactful relations vary by dataset and are typically content-related. For TUT2017, the top relations `is a variant of`, `has parent` and `scene occurs` pertain to acoustic scenes, including sound event variations, label hierarchy, and scene location.

Embedding Visualisations Appendix A.6 shows that the ZSAC performance varies across classes. To deeply understand why CLAP-KG improves ZSAC performance for certain classes while degrading it for others, we visualize the Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) projections of the embeddings, as shown in Figure 5. Although UMAP does not preserve exact distances, the resulting embedding clusters can still offer valuable insights into the relative data distribution. The top row of Figure 5 shows the mean audio embeddings (circle), the embedding of the top-1 CLAP predictions (star), and the top-1 CLAP-KG (triangles). Colors indicate different classes, with each subfigure using a distinct color scheme because of the different set of predictions. For each subfigure, we see multiple triangles as the CLAP predictions can be enriched by the KG in various ways depending on the tails. CLAP-KG enriches predictions when the ground-truth is *helicopter*, *bird chirping*, *crow*, *crackle*, and *cow*. These classes to which CLAP-KG brings the most improvement. Indeed, for all these classes, the CLAP-KG prediction clusters overlap with the audio embeddings, whereas the CLAP predictions remain disjoint.

To support a more balanced view, we also plot the 5 classes for which CLAP-KG degrades the performance, i.e. *cricket*, *rain*, *laughing*, *mouse click*, and *engine*, as shown by Figure 5 bottom. The audio and the correct CLAP predictions (the circle and star with the same color) indeed overlap, while sometimes CLAP-KG does not.

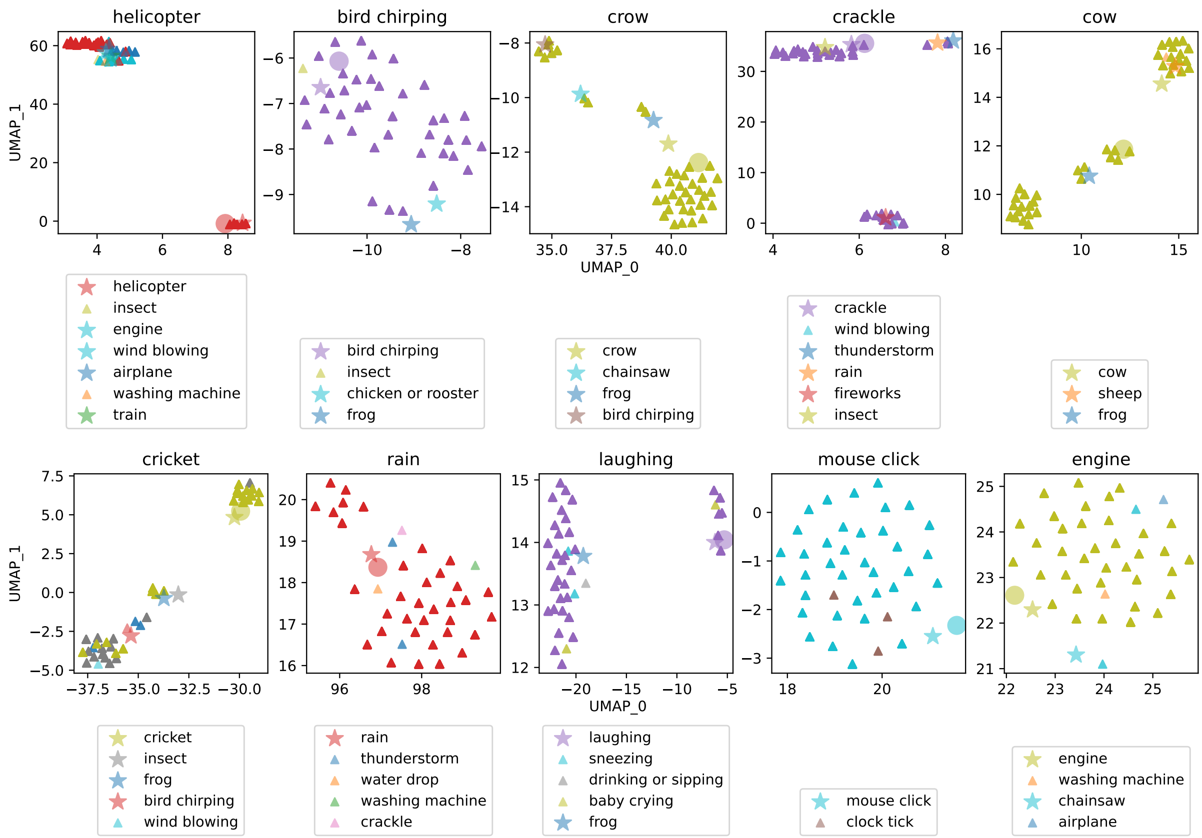

Figure 5: UMAP projection of the embeddings of CLAP audio (circle ●), top-1 CLAP prediction (star ★), and top-1 CLAP-KG predictions (triangle ▲). Colors indicate different classes, with each subfigure using a distinct color scheme. Top: the 5 classes rightmost in Figure 9 that CLAP-KG improves the performance. Bottom: the 5 classes leftmost in Figure 9 that CLAP-KG degrades the performance.

5.5 Discussion

Based on the observations and analysis above, we sum-up the following main findings:

A posthoc prediction recalibration with our AKG can boost ZSAC without further training or tuning. Note that in the proposed pipeline, the KG directly operates on CLAP predictions without further training.

Meaningful relations are key to integrating a KG due to the specificity of different datasets. As evidenced by Figure 4, relations that enhance the understanding of context and background knowledge of acoustic scenes augment the performance on TUT2017 by a large margin. This also points out that a powerful and generalizable KG must encompass a variety of relations.

Our AKG frees the efforts on prompt engineering and provides trackable reasoning. With the AKG, we can query audio-language models with only semantic cores, e.g. the class labels, eliminating the need for extensive prompt design. Furthermore, the KG predictions provide transparency into the classification process (through reasoning or factual knowledge retrieval), revealing both the predicted labels and their interrelations.

6 Conclusion

In this paper, we present `iKnow-audio`, a framework to enhance audio-language model predictions with knowledge graphs. We create the first audio knowledge graph (AKG) that encompasses rich relational semantics among everyday sounds. This structured knowledge is then encoded into a knowledge graph model to enhance predictions of an instantiated CLAP model. Our main finding is that instead of isolated semantic cores, AKG provides the necessary context and background knowledge for understanding sound events. The proposed method is post-hoc and lightweight, akin to Retrieval Augmented Generation (RAG), requiring neither fine-tuning nor prompt engineering when using audio-language models. It also holds potential for generalization to other tasks, such as question answering.

Limitations and Future Work

Despite the potential of the proposed method, we are aware of the following limitations of the current work and suggest the corresponding future directions: (1) **Shallow and Heuristic Reasoning**: Our approach currently performs only single-hop reasoning (tail prediction) over the knowledge graph (KG) and enriches prompts using simple string concatenation. This limits the depth and expressiveness of semantic inference. Future work could explore multi-hop reasoning as relations in the KG space can be chained. (2) **Noise and Incompleteness in the KG**: The KG was automatically constructed and cleaned, yet it may still contain noisy, generic, or missing triples. Additionally, link prediction from the knowledge embedding model can be unreliable for rare or ambiguous events, potentially introducing irrelevant or spurious concepts into the reasoning process. (3) **Limited Evaluation Scope**: We have not evaluated the method on music datasets, although the KG encodes music-related knowledge (through music-related labels from SALT). Extending evaluation to musical audio and broader domains would help assess the generality of the approach. (4) **Design and Efficiency Constraints**: The use of top-K selection for both CLAP and KG predictions may not capture the most informative evidence and could be biased toward frequent entities. Moreover, inference-time reasoning introduces additional computational overhead (through beam search). Future work may explore alternative sampling strategies and efficiency optimizations.

Acknowledgments

References

Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. 2021. PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. *Journal of Machine Learning Research*, (82):1–6.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. Fsd50k: An open dataset of human-labeled sound events.

Meng Gao, Yutao Xie, Wei Chen, Feng Zhang, Fei Ding, Tengjiao Wang, Jiahui Yao, Jiabin Zheng, and Kam-Fai Wong. 2025. Rerankgc: A cooperative retrieve-and-rerank framework for multi-modal knowledge graph completion. *Neural Networks*, page 107467.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Biao Gong, Shuai Tan, Yutong Feng, Xiaoying Xie, Yuyuan Li, Chaochao Chen, Kecheng Zheng, Yujun Shen, and Deli Zhao. 2024. Uknow: A unified knowledge protocol with multimodal knowledge graph datasets for reasoning and vision-language pre-training. *Advances in Neural Information Processing Systems*, 37:9612–9633.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, and 1 others. 2021. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37.

Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. 2024. Multimodal reasoning with multimodal knowledge graph. *arXiv preprint arXiv:2406.02030*.

Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. 2023. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36:13448–13466.

Hailun Lin, Yong Liu, Weiping Wang, Yinliang Yue, and Zheng Lin. 2017. Learning entity and relation embeddings for knowledge resolution. *Procedia Computer Science*, 108:345–354. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

Junming Liu, Siyuan Meng, Yanting Gao, Song Mao, Pinlong Cai, Guohang Yan, Yirong Chen, Zilin Bian, Botian Shi, and Ding Wang. 2025. Aligning vision to language: Text-free multimodal knowledge graph construction for enhanced llms reasoning. *arXiv preprint arXiv:2503.12972*.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen. 2017. DCASE 2017 challenge setup: Tasks, datasets and baseline system. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pages 85–92.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. Tut database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1128–1132.

Michel Olvera, Paraskevas Stamatiadis, and Slim Essid. 2024. A sound description: Exploring prompt templates and class descriptions to enhance zero-shot audio classification. In *The Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Abhirama Subramanyam Penamakuri, Kiran Chhatre, and Akshat Jain. 2025. Audiopedia: Audio qa with knowledge. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

J. Salamon, C. Jacoby, and J. P. Bello. 2014. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2017. Modeling relational data with graph convolutional networks.

Paraskevas Stamatiadis, Michel Olvera, and Slim Essid. 2024. Salt: Standardized audio event label taxonomy. *cities*, 17:26.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations (ICLR)*.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction.

Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. 2023. Tiva-kg: A multimodal knowledge graph with text, image, video and audio. In *Proceedings of the 31st ACM international conference on multimedia*, pages 2391–2399.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, pages 1112–1119. AAAI Press.

Xiaoyang Wei, Zografoula Vagena, Camille Kurtz, and Florence Cloppet. 2024. Integrating expert knowledge with vision-language model for medical image retrieval. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE.

A Appendix

A.1 Knowledge Graph Relation Schema

We define a schema comprising eight high-level relation categories, each reflecting a distinct aspect of auditory context. Each category includes a set of relations that guide the generation of plausible triples (head, relation, tail), where the head is a standardized sound event label (from SALT (Stamatiadis et al., 2024)) and the relation contextualizes its link to the tail concept. These categories are summarized in Table 3 and described as follows:

Co-occurrence and Temporal relations capture how sound events unfold over time or co-occur within sound scenes. Relations such as `co-occurs with`, `precedes`, `follows`, and `overlaps with` help model sequencing of events (e.g., "*thunder* `precedes` *lightning*").

Causal and Functional relations express underlying causes or functions of sound events, including `produces`, `caused by`, `triggers`, `indicates`, `responds to`, and `affects`. These relations allow the KG to represent inferential chains (e.g., "*siren* `triggers` *emergency response*") and explain sound occurrences based on physical or intentional causality.

Taxonomic and Hierarchical relations organize sounds into ontological structures using `is a type of`, `has subtype`, `is instance of`, `belongs to class`, and `is variant of`. These relations support reasoning about sound categories and enable class-based generalizations (e.g., "*laughter* `is a type of` *human sound*").

Spatio-Environmental Relations situate sound events within physical and environmental contexts through relations such as `occurs in`, `can be heard in`, `localized in`, `originates from`, and `associated with environment`. These are particularly valuable for acoustic scene classification and localization tasks.

Source and Agent Relations focus on the source of origin of a sound event. Relations like `emitted by`, `performed by`, `generated by`, `is sound of`, and `produced during` encode associations between sounds and their animate or inanimate sources (e.g., "*chirping* `performed by` *bird*").

Perceptual and Qualitative relations model human-centric interpretations of sound, using descriptors such as `has loudness`, `has pitch`, `has duration`, `has timbre`, `perceived as`, and `emotionally associated with`. These attributes provide complementary information that supports affective computing and perceptual modeling.

Modality-Crossing relations link auditory signals to language and vision, including `described by`, `associated with event`, `linked to visual`, and `transcribed as`. Such relations enable multimodal grounding and textual or visual alignment for sound events.

Intentionality relations express functional and normative expectations related to sound, via `invites action`, `used for`, `requires attention`, and `warns about`. These are particularly relevant for modeling listener responses and action-affording cues (e.g., "*doorbell* `invites action` *open door*").

Scene Composition and Event Structure captures how individual sound events compose or imply broader scenes or activities, through `part of scene`, `scene contains`, `event composed of`, `temporal component of`, and `entails event`. These relations provide a high-level abstraction of the acoustic scene and a structural prior for scene recognition.

A.2 Audio Knowledge Graph Statistics

In Figure 6 we present key statistics that provide a detailed characterization of the relational structure of the proposed knowledge graph. This includes measures of reflexivity, transitivity, and relation frequency distributions.

Total Relations, Heads and Tails summarize the volume and diversity of relational instances. The total relations count all occurrences, while unique heads and tails reflect the number of distinct entities appearing as the first (head) or second argument (tail) in each relation.

Reflexivity is evaluated by counting instances where the head and tail entities are identical. This highlights self-referential relations within the graph.

Transitivity is assessed by identifying triples where the relation can be inferred transitively (if (a, r, b) and (b, r, c) exist, then (a, r, c) is expected). The proportion of such inferred triples provides information on potential hierarchical or chain-like relational structures.

An overview of the global entity and relation counts, along with the 20 most frequent relations is summarized in Table 4.

A.3 Exemplary triples from the AKG

Table 5 presents a set of exemplary triples from the constructed knowledge graph. The first part of the table includes examples generated using a large language model (LLM), selected to depict a wide range of semantic relations such as causality, emotional association, perceptual attributes, and functional use. The second part provides examples derived from SALT, reflecting structured annotations grounded in taxonomies for everyday sound categorization. This combined presentation illustrates both the generative breadth of LLMs in synthetic data creation and the specificity of human-curated data, providing qualitative insight into the diverse relational structure captured in the graph.

A.4 CLAP-AKG Algorithm Description

Algorithm 1 details the full inference pipeline for knowledge-guided zero-shot audio classification using CLAP and a knowledge embedding model. Given an input audio sample and a set of candidate class labels, the algorithm first performs standard CLAP-based retrieval to identify the top-k most similar labels based on cosine similarity in the joint embedding space. For each top-ranked label, it queries a curated set of semantic relations $\mathcal{R}_q$ using the knowledge embedding model ϕ_{KG} to predict the most plausible tail entities. These tail entities are concatenated with the original label to form enriched, context-aware textual prompts. The CLAP text encoder then scores these prompts against the input audio. The final prediction is made by aggregating evidence from both the original and enriched prompts using a log-sum-exp fusion strategy, en-

Category	Example Relations	Purpose
Co-occurrence & Temporal	`co-occurs with`, `precedes`, `follows`, `overlaps with`	Capture temporal ordering and co-occurrence of sound events.
Causal & Functional	`produces`, `caused by`, `triggers`, `indicates`, `responds to`, `affects`	Encode causality, function, and event-response dynamics.
Taxonomic & Hierarchical	`is a type of`, `has subtype`, `is instance of`, `belongs to class`, `is variant of`	Structure sound events via type, class, and instance hierarchies.
Environmental	`occurs in`, `can be heard in`, `localized in`, `originates from`, `associated with environment`	Anchor sound events in physical, spatial, and environmental contexts.
Source & Agent	`emitted by`, `performed by`, `generated by`, `is sound of`, `produced during`	Link sounds to their generating sources.
Perceptual & Qualitative	`has loudness`, `has pitch`, `has duration`, `has timbre`, `perceived as`, `emotionally associated with`	Model perceptual properties and subjective qualities of sound.
Cross-modality	`described by`, `associated with event`, `linked to visual`, `transcribed as`	Establishes connections to textual or visual modalities.
Intentionality	`invites action`, `used for`, `requires attention`, `warns about`	Represent expectations, actions, or alerts invoked by sound.
Compositionality	`part of scene`, `scene contains`, `event_composed_of`, `temporal component of`, `entails event`	Capture hierarchical and compositional structure of scene and events.

Table 3: Relation schema for knowledge graph construction. Each category defines semantic relations that support rich contextualization of audio events.

abling semantic re-ranking of the top-k candidates. This procedure enhances both the interpretability and robustness of zero-shot classification by leveraging structured knowledge.

A.5 Prompt Templates for Triple Generation

To extract relational knowledge from large language models, we design a prompt template that guides the generation of plausible (head, relation, tail) triples grounded in sound event semantics. The prompt is tailored to elicit contextually relevant relations for each unique sound label in the SALT taxonomy. We apply it at scale to generate an initial pool of candidate triples, which are subsequently refined through a two-stage filtering process involving automated plausibility checks and manual curation. Figure 7 illustrates the prompt used for triple generation, while Figure 8 shows the prompt used to verify their semantic plausibility.

A.6 Additional Results

Per-class zero-shot audio classification performance In addition to the overall performance analysis in Section 5.4.2, we also investigate how CLAP-KG benefits individual classes. We consider ESC50 as an example and plot the class-wise classification performance of CLAP and CLAP-KG in. We notice that although the overall accuracy is increased by 2.2% as shown in Table 2, the class-wise performance varies. Large performance increase happens for *crow*, *crackle*, and *cow*, while CLAP-KG degrades performance for *cricket*, *rain*, and *laughing*.

A.7 Dataset Licenses

For transparency, we provide a comprehensive summary of the licensing terms associated with each dataset used in our experiments in Table 6. All datasets are publicly available and widely used in

Knowledge Graph Summary					
	Subset	**Triples**	**Relations**	**Heads**	**Tails**
Overall Stats	**Clean**	18,348	47	857	4,282
	Noisy	49,215	47	860	11,063
	Test	2,039	46	673	1,068

Top 20 Most Frequent Relations (Split by Clean and Noisy Sets)							
#	**Relation**	**Triples**		**Heads**		**Tails**	
		Clean	**Noisy**	**Clean**	**Noisy**	**Clean**	**Noisy**
1	`has subtype`	2552	3773	331	528	1020	1731
2	`belongs to class`	2242	2739	828	835	252	471
3	`occurs in`	2052	2982	550	622	347	640
4	`has children`	907	907	211	211	773	773
5	`has sibling`	890	890	760	760	207	207
6	`has parent`	886	886	764	764	206	206
7	`can be heard in`	631	1212	289	366	249	378
8	`localized in`	623	893	226	241	268	355
9	`part of scene`	564	1531	164	253	337	752
10	`is a type of`	529	929	233	304	251	460
11	`generated by`	501	936	255	327	277	450
12	`described by`	393	661	242	295	368	627
13	`event composed of`	390	1368	236	441	284	877
14	`produced during`	363	712	161	219	241	395
15	`overlaps with`	348	2009	185	434	237	844
16	`associated with environment`	330	593	128	180	210	323
17	`precedes`	308	1010	122	227	220	643
18	`originates from`	304	579	138	172	207	377
19	`warns about`	272	1854	97	353	187	854
20	`emitted by`	254	319	135	149	149	183

Table 4: Summary statistics for the knowledge graph. The upper section presents overall statistics including the number of triples, relations, head and tail entities. The lower section lists the 20 most frequent relations, split by clean and noisy subsets, with counts of associated triples, heads, and tails.

academic research on environmental sound classification.

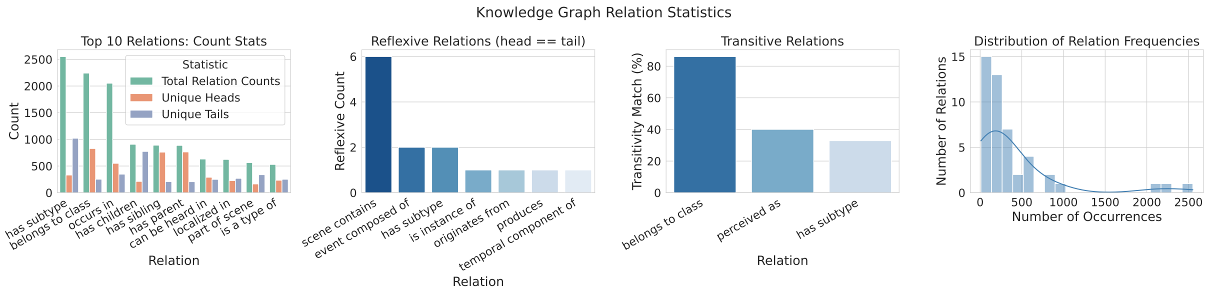

Figure 6: Overview of key statistics for relations of the clean set in the knowledge graph. **(a)**: Distribution of counts, unique heads, and unique tails for the top 10 most frequent relations. **(b)**: Counts of reflexive relations where the head equals the tail. **(c)**: Proportion of transitive triples identified among the total triples per relation. **(d)**: Distribution of relation frequencies.

"You are an expert in sound event classification and knowledge graph generation. Given a sound event label, your task is to reason about and, if appropriate, generate knowledge graph triples that describe real-world, common-sense relationships between the sound event and other entities or events. The relation type is: {relation_type}. The relation details are: {relation_details}. Here is an example for guidance: {examples}.

Step 1: Reason about the plausibility of generating real-world, common-sense triples for the sound event label: {label_name}, using the relation type:{relation_type}. Determine if this type of relation is meaningfully applicable to the event in a way that reflects actual, observable relationships in the world.
If the relation type is not applicable or would lead to speculative, forced, or non-sensical triples, conclude that no valid triples can be generated.

Step 2: If the relation is applicable and meaningful, generate a list of plausible, real-world triples grounded in common sense. Ensure that each triple reflects knowledge that a reasonable person would accept as true in everyday understanding.
There is no fixed number of triples required, but include only those that are relevant, accurate, and justifiable by common sense.
Respond with only the final list of triples in the exact format: [[head1, relation, tail1], [head2, relation, tail2], ...].
If in Step 1 you determine that no meaningful triples can be generated, respond with an empty list: [].
Do not include any reasoning or explanation in the final output. The head should strictly be the label name: {label_name}."

Figure 7: Prompt template to generate synthetic triples via LLM.

"You are an expert in knowledge graphs for audio understanding. Given a triple in the format [head, relation, tail], assess whether it is pertinent for inclusion in a knowledge graph for audio understanding. The head represents a sound event label, i.e., a sound or an abstraction of the sound emitted, implied, or perceptually associated with an entity. A triple is pertinent if it is non-speculative, grounded in common-sense and real-world experience, and contributes to a taxonomical, hierarchical, temporal, causal, perceptual, compositional, or phisical contextual understanding of sound events. Reject triples which are vague, speculative, or not useful for structuring knowledge about sound. Is the triple {kg_triple} pertinent to structure knowledge about sound? Answer strictly "Yes" or "No" without any reasoning or explanation in the final output."

Figure 8: Prompt template to verify synthetic triples via LLM.

	SALT Label	Head	Relation	Tail
Triple examples (generated by LLM)				
1	*vehicle engine*	*vehicle engine*	caused by	*combustion*
2	*chicken crowing*	*chicken crowing*	caused by	*rooster*
3	*smoke alarm*	*smoke alarm*	caused by	*smoke*
4	*crying*	*crying*	emotionally associated with	*sadness*
5	*cello*	*cello*	emotionally associated with	*melancholy*
6	*lullaby*	*lullaby*	emotionally associated with	*calmness*
7	*coffee machine*	*coffee machine*	has duration	*medium*
8	*timpani*	*timpani*	has duration	*long*
9	*cap gun*	*cap gun*	has duration	*short*
10	*bird*	*bird*	has pitch	*high*
11	*humming*	*humming*	has pitch	*low*
12	*flute*	*flute*	has pitch	*high*
13	*thunderstorm*	*thunderstorm*	indicates	*thunder*
14	*marching*	*marching*	indicates	*parade*
15	*firecracker*	*firecracker*	indicates	*celebration*
16	*maraca*	*maraca*	is instance of	*percussion instrument*
17	*giggling*	*giggling*	is instance of	*laughter*
18	*microphone*	*microphone*	is instance of	*audio recording device*
19	*fireworks*	*fireworks*	perceived as	*celebratory*
20	*castanets*	*castanets*	perceived as	*rhythmic instrument*
21	*pulse*	*pulse*	perceived as	*heartbeat rate*
22	*flute*	*flute*	performed by	*orchestra*
23	*kwaito music*	*kwaito music*	performed by	*musicians*
24	*playing guitar*	*playing guitar*	performed by	*guitarist*
25	*clock tick*	*clock tick*	precedes	*door opening*
26	*electric guitar*	*electric guitar*	precedes	*composing music*
27	*dog*	*dog*	precedes	*yelping*
28	*mantra*	*mantra*	used for	*self-improvement*
29	*whistle*	*whistle*	used for	*alerting*
30	*knife*	*knife*	used for	*self-defense*
Triple examples (derived by SALT)				
31	*pigeon dove*	*pigeon dove*	belongs to class	*bird*
32	*large rotating saw*	*large rotating saw*	belongs to class	*sawing*
33	*vehicle compressor*	*vehicle compressor*	belongs to class	*large vehicle*
34	*speech*	*speech*	has children	*chatter*
35	*wild animal*	*wild animal*	has children	*roar*
36	*bowed string instrument*	*bowed string instrument*	has children	*cello*
37	*whoosh swoosh swish*	*whoosh swoosh swish*	has parent	*wind*
38	*bouncing on trampoline*	*bouncing on trampoline*	has parent	*jumping*
39	*swimming*	*swimming*	has parent	*water activity*
40	*swimming*	*swimming*	has sibling	*diving*
41	*whoosh swoosh swish*	*whoosh swoosh swish*	has sibling	*rustling*
42	*bouncing on trampoline*	*bouncing on trampoline*	has sibling	*bouncing ball*
43	*piano*	*piano*	has subtype	*grand piano*
44	*music genre*	*music genre*	has subtype	*jazz*
45	*vehicle*	*vehicle*	has subtype	*bicycle*
46	*smash or crash*	*smash or crash*	occurs in	*kitchen*
47	*drum kit*	*drum kit*	occurs in	*train station*
48	*clatter*	*clatter*	occurs in	*gym*

Table 5: Representative examples of knowledge graph triples. The first section includes examples generated using a large language model (LLM), grouped by semantic relation types such as causality, perception, and functionality. The second section includes examples extracted from the SALT. Both sets illustrate complementary richness and diversity of relation types from automated and curated construction approaches.

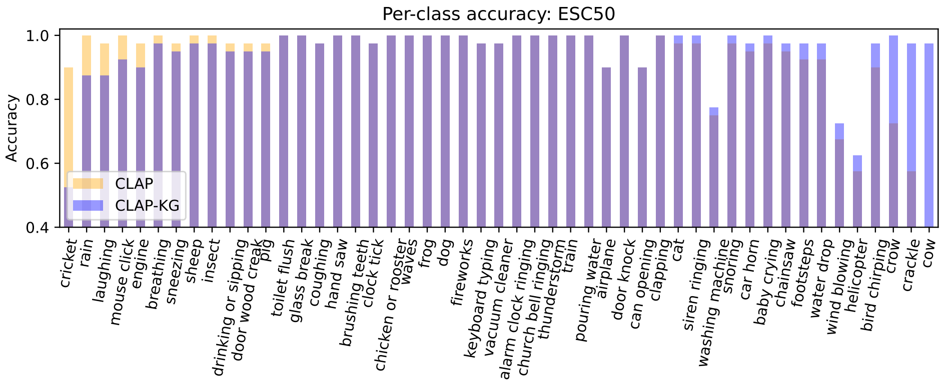

Figure 9: Per-class zero-shot audio classification accuracy with CLAP and CLAP-KG on ESC50 dataset.

Algorithm 1 Knowledge-Guided CLAP Inference

Require: Input audio $a \in \mathcal{A}$, label set $C = \{c_1, \ldots, c_N\} \subset \mathcal{L}$, CLAP encoders $\phi_{\text{A}}, \phi_{\text{T}}$, knowledge embedding model ϕ_{KG}, relation set $\mathcal{R}_q \subset \mathcal{R}$, top-$k$ parameters k, m
Ensure: Predicted label $\tilde{c} \in C$

Encode audio: $\mathbf{a} \leftarrow \phi_{\text{A}}(a)$
Encode labels: $\mathbf{c}_i \leftarrow \phi_{\text{T}}(c_i)$ for all $c_i \in C$
Compute similarities: $s(c_i) \leftarrow \text{sim}(\mathbf{a}, \mathbf{c}_i)$
Retrieve top-k labels: $C_k = \{c^{(1)}, \ldots, c^{(k)}\} \leftarrow \texttt{TopK}(\{s(c_i)\}, k)$
Initialize enriched prompt set: $\mathcal{P} \leftarrow \emptyset$
for all $c \in C_k$ **do**
    for all $r \in \mathcal{R}_q$ **do**
        Predict top-m tails: $\mathcal{T}_c^r \leftarrow \texttt{TopM}(\phi_{\text{KG}}(c, r, \cdot), m)$
        for all $t \in \mathcal{T}_c^r$ **do**
            Form enriched prompt: $p_{c,t} \leftarrow \texttt{concat}(c, t)$
            Add $p_{c,t}$ to $\mathcal{P}$
        end for
    end for
end for
Encode enriched prompts: $\mathbf{p}_j \leftarrow \phi_{\text{T}}(p_j)$ for all $p_j \in \mathcal{P}$
Compute prompt similarities: $s(p_j) \leftarrow \text{sim}(\mathbf{a}, \mathbf{p}_j)$
for all $c \in C_k$ **do**
    Retrieve prompt scores: $\{s(p_j) \mid p_j \in P_c\}$
    Aggregate score: $\tilde{s}(c) \leftarrow \log\left(\exp(s(c)) + \sum_{p_j \in P_c} \exp(s(p_j))\right)$
end for
Predict final label: $\tilde{c} \leftarrow \arg\max_{c \in C_k} \tilde{s}(c)$
return $\tilde{c}$

Dataset	License
ESC50 (Piczak)	CC BY-NC 3.0 (Attribution-NonCommercial)
UrbanSound8K (Salamon et al., 2014)	CC BY-NC 3.0 (Attribution-NonCommercial)
TUT2017 (Mesaros et al., 2016)	Custom EULA: Non-commercial scientific use only
FSD50K (Fonseca et al., 2022)	CC BY 4.0 (Attribution)
AudioSet (dataset) (Gemmeke et al., 2017)	CC BY 4.0 (Attribution)
AudioSet (ontology) (Gemmeke et al., 2017)	CC BY-SA 4.0 (Attribution-ShareAlike)
DCASE17-T4 (Mesaros et al., 2017)	Follows AudioSet licensing

Table 6: Summary of dataset licenses used in this study.