# Resonance: Open-Ended Instruction-Based Audio Retrieval

Anonymous ACL submission


## Abstract

Audio retrieval – the task of finding relevant audio given a reference clip, involves diverse and nuanced search intents that cannot be captured by surface-level acoustic similarity alone. Existing approaches are limited in that they operate purely on high-level audio semantics and characteristics, without supporting custom or open-ended retrieval criteria. In this paper, we introduce **Resonance**, a novel framework for instruction-guided audio retrieval that allows users to specify flexible search intents via open-ended and free-form text instructions in addition to an audio query. Resonance leverages both audio and text encoders and introduces a unified multi-head attention pooler that fuses the multi-modal inputs into a single embedding for retrieval. To enable training and evaluation, we present **Resonance-Set**, a large-scale dataset of 350K training and 1.5K evaluation triplets, constructed through a novel five-stage curation pipeline. Experimental results show that Resonance achieves 24.5% to 62.2% improvements over the baselines. We will release our code, models, and datasets.


## 1 Introduction

Content-based retrieval, or query by example, is the task of retrieving media samples from a database that are similar to a user-provided query. It is a long-standing problem in deep learning with broad applications across modalities – including audio, vision, and text – for tasks such as recommendation (Covington et al., 2016), search (Hershey et al., 2017), content creation (Liang et al., 2024), etc. In this work, we specifically focus on audio, defined as non-speech and non-verbal sounds, and music.

Humans have the remarkable ability to detect and relate novel sound events – or acoustic objects – based on how they stand out against a background (McDermott and Simoncelli, 2011). This capacity to generalize similar sounds beyond exact matches is crucial for navigating the world, and is

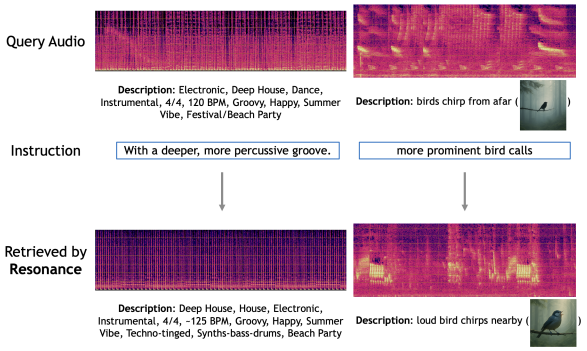


Figure 1: Illustration of our proposed task, *open-ended instruction-based audio retrieval*. Given a query audio and a free-form instruction that specifies a modification, the system retrieves an audio sample that follows the instruction while preserving other properties (images and descriptions are included for readability and illustrative purposes only).

equally important for AI agents and robots. While semantic similarity and retrieval have been extensively explored in images and video (Yoon et al., 2021; Liu et al., 2019), audio-to-audio retrieval has received comparatively limited attention (Manocha et al., 2018). Prior work has primarily focused on music information retrieval (Komatsu et al., 2025; Wu et al., 2025) or text-to-audio retrieval (Wu* et al., 2023; Ghosh et al., 2023), leaving open challenges in general-purpose, semantics-driven audio-to-audio retrieval.

We argue this gap arises because real-world audio encapsulates complex, multifaceted semantics that make "similarity" inherently ambiguous. Two audio recordings may share some perceptual traits while differing in others, or differ in surface form while conveying overlapping semantic properties. Consequently, for audio search, simply retrieving items based on acoustic content similarity is insufficient. Instead, as in image search (Yoon et al., 2021), it is often more natural to pair an example audio with an explicit search intent describing the aspects that should be emphasized or varied. For example, when querying with a song, a user might want "a more rhythmic version," "a darker, ambi-

ent rendition," or "a track with stronger vocals." This motivates incorporating natural language instructions alongside the query audio to articulate open-ended search intents. Such instructions are indispensable for capturing the richness of user needs and guiding retrieval toward semantically appropriate results. Ultimately, models should be able to jointly interpret audio and language, grounding retrieval in both the acoustic content and the intent expressed in free-form text.

**Main Contributions.** In this paper, we introduce **Resonance**, a simple yet effective method for *open-ended instruction-based audio-to-audio retrieval*. The task is defined as follows: given a query audio and a user intent in free-form natural language instruction, the goal is to retrieve an audio sample from a database that best satisfies the specified intent. We propose a simple architecture that combines an audio encoder and a text encoder, and introduces a unified multi-head attention pooler that fuses the multimodal inputs into a single embedding used for retrieval. Since this task is novel, we also construct **Resonance-Set**, a large-scale, high-quality dataset comprising 350K training and 1.5K evaluation triplets, created through a five-stage curation pipeline. Our model is trained with a simple contrastive objective. To the best of our knowledge, **Resonance** is the first method to support such open-ended instruction-driven audio retrieval. In summary, our contributions are:

1. We introduce the task of open-ended instruction-based audio-to-audio retrieval.
2. We propose **Resonance**, a simple architecture that integrates audio and text encoders with a multi-head attention pooler for this task.
3. We curate **Resonance-Set**, a large-scale dataset with 350K training and 1.5K evaluation triplets, constructed through a novel five-stage pipeline for this task.
4. We demonstrate that **Resonance** outperforms strong baselines by 24.5% to 62.2% on **Resonance-Set-Test**.

## 2 Related Work

CLIP (Radford et al., 2021) introduced a scalable framework for multimodal retrieval, learning a shared vision–language embedding space through a dual-encoder architecture and contrastive training. By aligning text and images in the same space, CLIP enabled flexible retrieval across modalities (image–image, image–text, text–image). Inspired by this success, CLAP (Elizalde et al., 2023) extended the contrastive dual-encoder paradigm to the audio–language domain, achieving state-of-the-art performance on tasks such as text–audio retrieval and zero-shot classification. Subsequent research has largely focused on improving CLAP representations (Wu* et al., 2023; Ghosh et al., 2023, 2025). Related approaches include Wav2CLIP (Wu et al., 2022), AudioCLIP (Guzhov et al., 2022), and CoLLAT (Silva et al., 2023). However, prior work has primarily addressed text–audio or audio–text retrieval and zero-shot classification, while largely overlooking the setting of *audio-to-audio retrieval with open-ended search intents*. A few early efforts have considered similarity-based audio-to-audio retrieval (Manocha et al., 2018), but these methods do not incorporate natural language instructions to capture flexible user intents.

Open-ended instruction-based retrieval was introduced by UniIR (Wei et al., 2024) in the vision domain, allowing an extra text instruction to flexibly define the retrieving request. For example, one may retrieve an image of a news article with the header of a particular style. This is achieved by curating diverse multimodal retrieval triplets (query, instruction, output) and fusing the instruction into the query when computing cosine similarity. Follow-up research further improved retrieval quality by scaling data and researching different architectures (Zhang et al., 2024; Lin et al., 2025). However, open-ended instruction-based retrieval is less studied in the audio domain mainly due to a lack of training data, well-defined retrieval tasks, and variable (potentially very long) audio lengths. To our knowledge, this work presents the first open-ended instruction-based retrieval model for audio.

## 3 Methodology

### 3.1 Resonance-Set Data Construction

To the best of our knowledge, we are the first to formally explore the task of open-ended instruction-based audio retrieval, and no existing public datasets are available. To address this gap, we curate **Resonance-Set**, a large-scale dataset constructed via a novel five-step pipeline. **(1) Audio Collection.** We compile a diverse pool of audio from 10 existing datasets. For general sounds, we use AudioSet (Gemmeke et al., 2017) (including AudioSet-Strong (Hershey et al., 2021) and AudioCaps (Kim et al., 2019)), Clotho-v2 (Drossos

et al., 2020), CochlScene (Jeong and Park, 2022), YouTube-8M (Abu-El-Haija et al., 2016), SoundDescs (Koepke et al., 2022), MiraData (Ju et al., 2024), and ReCap (Nguyen et al., 2025). For music, we include Million Song Dataset (MSD) (Dataset, 2011), and Music4All (Santana et al., 2020). **(2) Audio Captioning.** Each audio is captioned with Audio Flamingo 3 (Goel et al., 2025), producing detailed descriptions that capture sound events, musical attributes, and spoken elements. **(3) Audio Clustering.** We embed each audio into $\mathbb{R}^{768}$ and apply Faiss (Douze et al., 2024) clustering. For every query audio $A$, we identify its top-$n = 2$ most similar ($A^p$) and most dissimilar ($A^n$) audios. **(4) Metadata Curation.** For each audio, we compute a rich set of metadata to guide instruction generation. This includes tags, duration, loudness, pitch, reverb, noise, brightness, and fade following Kumar et al. (2024). For music, we additionally compute emotion labels, key, tempo (BPM), and chord progressions. A full list of metadata and tools is provided in Appendix B. **(5) Instruction Generation.** Finally, we prompt GPT-5 with the metadata and pairs $(A, A^p)$ and $(A, A^n)$ to generate natural language instructions describing how $A^p$ or $A^n$ relates to $A$. The instructions span diverse categories, including mood, cultural or era-specific context, production style, structural differences, etc, for music, and loudness, quality, background noise variations, etc., for environmental sounds. We keep our instructions short, within 5-10 words, and do not generate complex instructions with multiple intents. The prompting strategy is illustrated in Fig. 3 and Fig. 4, with representative dataset examples shown in Table 4. **(6) Quality Filtering.** After constructing the dataset, we perform an additional filtering step to remove noisy instructions and incorrect pairs. Using GPT-4o-audio (Hurst et al., 2024), we provide the individual audios along with the generated instruction and ask the model to assess whether the instruction is plausible.

**Resonance-Set-Test.** To build a high-quality evaluation set, we uniformly sample ≈1% of the generated data across categories and have audio professionals manually validate the instructions with respect to each audio pair. The final dataset comprises 350K training and 1.5K evaluation triplets.

### 3.2 Resonance Model

**Architecture.** As shown in Fig. 2, we adopt a dual-encoder architecture, with shared parameters (Zhang et al., 2024). Both the audio and text encoders are initialized from a custom CLAP model, pre-trained at scale, with AF-Whisper (Goel et al., 2025) as the audio and T5-large (Raffel et al., 2020) as the language backbone. We choose AF-Whisper over HTS-AT (Chen et al., 2022), as it provides denser and richer representations, and find that CLAP-style joint pre-training further improves alignment between modalities. To enable the integration of both modalities, we add multiple self-attention layers and introduce a multi-head attention pooler that compresses the multimodal inputs into a single embedding $r$ for retrieval. Since the retrieval targets consist only of audio without accompanying text, we append an empty string (“”) as the text input, thereby transforming each target into a multimodal pair. We denote the resulting embeddings as $r_q$ for the query (audio$_q$, text) and $r_t$ for the target (audio$_t$, “”).

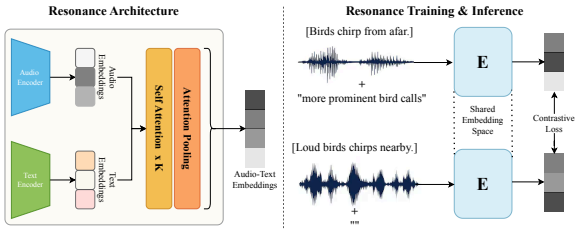


Figure 2: Overview of the Resonance architecture, training, and inference. The model encodes an audio query and an instruction into a joint embedding, and aligns query–target pairs in a shared embedding space via a contrastive loss.

**Training.** To train **Resonance**, we adopt a contrastive learning objective inspired by CLIP and CLAP. A key challenge is that the query audio itself (audio$_q$) can serve as a strong confounder for the multimodal query pair (audio$_q$, text). To mitigate this, inspired by Zhang et al. (2024), we introduce an additional hard negative by encoding the pair (audio$_q$, “”), denoted as $r'_t$.

To enrich the pool of negatives, for each query, we include both these query-specific negatives and all target negatives within the batch. The loss for the $i$-th training instance is then defined as:

$$\mathcal{L}_i = -\log \frac{\exp\left(\text{sim}(r_q^i, r_t^i)/\tau\right)}{\sum_{j=1}^{N}\Big[\exp\left(\text{sim}(r_q^i, r_t^j)/\tau\right) + \exp\left(\text{sim}(r_q^i, r_t'^j)/\tau\right)\Big]} \quad (1)$$

where sim(·) denotes cosine similarity

$$\text{sim}(r_q, r_t) = \frac{r_q^\top r_t}{\|r_q\|\,\|r_t\|},$$

$N$ is the batch size, and $\tau$ is a temperature parameter for logit scaling.

## 4 Experimental Setup

**Hyper-parameters.** We train **Resonance** with a learning rate of 5e-4 and an effective batch size (BS) of 256 for 10 epochs. We use 8 NVIDIA A100 80GB GPUs for training. For CLAP pre-training, we use the same hyperparameters and the original AF-Whisper training data (Goel et al., 2025). The final model is ≈1.5B parameters.

**Baselines and Ablations.** As this is the first work to study instruction-guided audio-to-audio retrieval, we design a set of baselines and ablations to evaluate Resonance. **Baselines.** 1. **Random:** A simple baseline that retrieves a random audio from the database. 2. **CLAP:** We evaluate several existing CLAP variants (including our re-trained version). For each query, we obtain embeddings from the audio and text encoders and combine them via either (i) **concatenation** (along the embedding dimension) or (ii) **mean pooling**. 3. **Audio-only:** Retrieval is performed using only the query audio embedding. 4. **Text-only:** Retrieval is performed using only the text instruction embedding. **Ablations.** 1. **w/ Mean Pooling:** Instead of attention pooling, we apply mean pooling over embeddings. 2. **w/o Self-Attention:** We remove the self-attention layers and apply attention pooling directly on concatenated audio and text embeddings across the sequence dimension. 3. **w/o CLAP Pre-training:** We fine-tune Resonance on Resonance-Set without initializing from CLAP pre-trained weights. 4. **w/o Instruction:** Retrieval is performed using only the query audio, ignoring the instruction input. All results have been averaged across 3 runs.

## 5 Results and Analysis

**Quantitative.** Table 1 presents a comparison between Resonance and the proposed baselines. We evaluate all models on **Resonance-Set-Test** using three standard retrieval metrics – R@1, R@5, and R@10. **Resonance** achieves substantial gains over CLAP-style baselines, highlighting the importance of task-specific fine-tuning. Furthermore, our ablation results demonstrate the contribution of the self-attention layers and the attention pooling module, validating the effectiveness of our architectural design choices.

**Qualitative.** Table 2 presents qualitative examples illustrating both the strengths and limitations of

| Method | R@1 | R@5 | R@10 |
|---|---|---|---|
| Random | 0.07 | 0.33 | 0.67 |
| Audio-only | 16.6 | 19.0 | 35.4 |
| Text-only | 4.2 | 11.4 | 27.8 |
| LAION-CLAP | 28.4 | 45.2 | 62.4 |
| MS-CLAP | 26.0 | 40.4 | 60.8 |
| ReCLAP | 35.6 | 52.0 | 69.0 |
| Resonance *(ours)* | **64.8** | **76.4** | **89.0** |
| w/ Mean Pooling | 60.2 | 71.0 | 79.4 |
| w/o Self-Attention | 55.6 | 68.4 | 75.2 |
| w/o CLAP Pre-training | 58.0 | 65.4 | 73.2 |
| w/o Instruction | 10.2 | 22.2 | 28.0 |
| w/ HTS-AT | 49.0 | 55.2 | 63.8 |

Table 1: Comparison of model performance on Resonance-Set-Test with baselines. Grey rows indicate ablations.

**Resonance**. As shown, Resonance struggles in cases where the instruction involves understanding spoken content (e.g., "casual conversations") or expresses multiple intents within a single query. In contrast, even the strongest CLAP-based models fail in most scenarios, except when clear and explicit cues are present in the instruction itself – highlighting the inherent complexity of open-ended, instruction-based audio retrieval.

| Instruction | ReCLAP | Resonance |
|---|---|---|
| Higher density of rapid, high-pitched chirps. | ✗ | ✓ |
| A more intense and higher-pitched electrical hum accompanied by a sharp, percussive onset. | ✗ | ✗ |
| fewer distinct high-frequency sounds, more consistent low-frequency hum | ✗ | ✗ |
| More casual conversations | ✗ | ✗ |
| A rock-infused version | ✗ | ✓ |
| More melancholic and string-heavy musical backing. | ✗ | ✓ |

Table 2: Performance comparison between ReCLAP (CLAP-based) and Resonance on instances from Resonance-Set-Test.

## 6 Conclusion

We propose Resonance, a simple yet effective framework for instruction-guided audio retrieval, which enables users to specify flexible search intents through free-form text instructions alongside an audio query. To support this task, we curate a new large-scale dataset using a five-stage pipeline. Resonance achieves significant improvements over competitive baselines, and we believe this new task will open up promising directions for advancing semantic and intent-driven audio retrieval.

## Limitations and Future Work

As part of future work, we would like to address the current limitations of Resonance, including:

1. Resonance currently handles only single-intent, relatively simple instructions. Extending the framework to multi-intent, compositional, and longer instructions is an important next step.

2. Our dataset contains only English instructions. Future work should expand Resonance-Set to multiple languages to support multilingual and cross-lingual retrieval.

3. Resonance is not optimized for cases where the instruction itself contains spoken language content, as such examples are absent in training. Incorporating these scenarios remains an open direction.

## References


Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. Htsat: A hierarchical token-semantic audio transformer for sound classification and detection. In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP.

Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In Proceedings of the 10th ACM conference on recommender systems, pages 191–198.

Million Song Dataset. 2011. Million song dataset.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. arXiv preprint arXiv:2401.08281.

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

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.

Jort F Gemmeke, Daniel PW 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. IEEE.

Sreyan Ghosh, Sonal Kumar, Chandra Kiran Reddy Evuru, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2025. Reclap: Improving zero shot audio classification by describing sounds. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Evuru, S Ramaneswaran, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2023. Compa: Addressing the gap in compositional reasoning in audio-language models. arXiv preprint arXiv:2310.08753.

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

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 976–980. IEEE.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. Cnn architectures for large-scale audio classification. In 2017 ieee international conference on acoustics, speech and signal processing (icassp), pages 131–135. IEEE.

Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. 2021. The benefit of temporally-strong labels in audio event classification. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 366–370. IEEE.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276.

Il-Young Jeong and Jeongsoo Park. 2022. Cochlscene: Acquisition of acoustic scene data using crowdsourcing. In 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pages 17–21. IEEE.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu,

and Ying Shan. 2024. Miradata: A large-scale video dataset with long durations and structured captions. Advances in Neural Information Processing Systems, 37:48955–48970.

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

A Sophia Koepke, Andreea-Maria Oncescu, João F Henriques, Zeynep Akata, and Samuel Albanie. 2022. Audio retrieval with natural language queries: A benchmark study. IEEE Transactions on Multimedia, 25:2675–2685.

Tatsuya Komatsu, Hokuto Munakata, Takuya Hasumi, and Yusuke Fujita. 2025. Aligned contrastive learning for text-to-music retrieval. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE.

Sonal Kumar, Prem Seetharaman, Justin Salamon, Dinesh Manocha, and Oriol Nieto. 2024. Sila: Signal-to-language augmentation for enhanced control in text-to-audio generation. arXiv preprint arXiv:2412.09789.

Jinhua Liang, Huan Zhang, Haohe Liu, Yin Cao, Qiuqiang Kong, Xubo Liu, Wenwu Wang, Mark D Plumbley, Huy Phan, and Emmanouil Benetos. 2024. Wavcraft: Audio editing and generation with large language models. arXiv preprint arXiv:2403.09527.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. MM-EMBED: UNIVERSAL MULTIMODAL RETRIEVAL WITH MULTIMODAL LLMS. In The Thirteenth International Conference on Learning Representations.

Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. arXiv preprint arXiv:1907.13487.

Pranay Manocha, Rohan Badlani, Anurag Kumar, Ankit Shah, Benjamin Elizalde, and Bhiksha Raj. 2018. Content-based representations of audio using siamese neural networks. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3136–3140. IEEE.

Josh H McDermott and Eero P Simoncelli. 2011. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron, 71(5):926–940.

Thinh-Phuc Nguyen, Thanh-Hai Nguyen, Gia-Huy Dinh, Lam-Huy Nguyen, Minh-Triet Tran, and Trung-Nghia Le. 2025. Recap: Event-aware image captioning with article retrieval and semantic gaussian normalization. arXiv preprint arXiv:2509.01259.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140):1–67.

Igor André Pegoraro Santana, Fabio Pinhelli, Juliano Donini, Leonardo Catharin, Rafael Biazus Mangolin, Valéria Delisandra Feltrim, Marcos Aurélio Domingues, et al. 2020. Music4all: A new music database and its applications. In 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), pages 399–404. IEEE.

Dadallage AR Silva, Spencer Whitehead, Christopher Lengerich, and Hugh Leather. 2023. Collat: On adding fine-grained audio understanding to language models using token-level locked-language tuning. Advances in Neural Information Processing Systems, 36:63197–63209.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In European Conference on Computer Vision, pages 387–404. Springer.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2clip: Learning robust audio representations from clip. In ICASSP 2022.

Junda Wu, Zachary Novack, Amit Namburi, Hao-Wen Dong, Carol Chen, Jiaheng Dai, and Julian McAuley. 2025. Futga-mir: Enhancing fine-grained and temporally-aware music understanding with music information retrieval. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE.

Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP.

Sangwoong Yoon, Woo Young Kang, Sungwook Jeon, SeongEun Lee, Changjin Han, Jonghun Park, and Eun-Sol Kim. 2021. Image-to-image retrieval by learning similarity between scene graphs. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 10718–10726.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. 2024. MagicLens: Self-supervised image retrieval with open-ended instructions. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pages 59403–59420. PMLR.

# A Appendix

In the Appendix, we provide:

1. Section B: Descriptors & Tools
2. Section C: Baseline Details
3. Section D: Prompts
4. Section E: Annotation Guidelines
5. Section F Attribute Examples
6. Section G: Dataset Licenses
7. Section H: Use of AI
8. Section I: Broader Impact & Risks

# B Descriptors & Tools

In Table 3, we list all the tools and libraries that we used to extract information for each audio like - reverb noise, brightness, pitch and duration.

| Descriptor | Tool Used |
|---|---|
| Reverb | Timbral Models[1] |
| Noise | Torchaudio[2] |
| Brightness | Torchaudio |
| Pitch | Torchcrepe |
| Duration | Torchaudio |
| Fade | Proprietary |
| Beats | Madmom[3] |
| Chords | Essentia[4] |

Table 3: Tools & libraries used to extract the DSP features from audios.

# C Baselines

**LAION-CLAP** (Wu* et al., 2023) proposes a large-scale contrastive language-audio pretraining model that leverages a newly introduced dataset called LAION-Audio-630K, which includes over 630k audio-text pairs. The model combines audio and text encoders with feature fusion and keyword-to-caption augmentation, improving performance on text-to-audio retrieval, zero-shot audio classification, and supervised audio classification tasks.

**ReCLAP** (Ghosh et al., 2025) builds on the work of LAION-CLAP, and introduces an enhanced CLAP model trained with rewritten audio captions to improve zero-shot audio classification (ZSAC) and retrieval tasks. The ReCLAP model is trained on ≈2.3M audio-caption pairs.

**MS-CLAP** (Elizalde et al., 2023) is an audio language model trained with contrastive learning between audio data and their corresponding natural language descriptions. It extracts representations from both audio and text encoders.

# D Prompts

Fig. 4 and Fig. 4 shows our prompting strategy to generate instructions. We keep the instructions short and concise.

# E Annotation Guidelines for Resonance-Set

The goal of annotation in **Resonance-Set** is to ensure high-quality evaluation data for audio–instruction alignment. Annotators verify whether automatically generated instructions are plausible, accurate, and concise descriptions of the relation between audio pairs. Each task involves listening to audio triplets $(A, A^p, A^n)$ along with the generated instruction. The annotator assesses if the instruction is semantically and perceptually valid for the given pair.

## E.1 Annotator Hiring & Requirements

- **Background:** Audio professionals (musicians, sound engineers, researchers) are preferred. Annotators with prior crowdsourcing/audio labeling experience may also be considered.
- **Skills:** Ability to distinguish common sound events (speech, footsteps, car horns, applause) and, for music, familiarity with features such as tempo, pitch, mood, and instrumentation.
- **Equipment:** High-quality headphones in a quiet environment; reliable internet connection.
- **Training:** Annotators undergo a short tutorial with practice examples and gold-standard checks before formal annotation begins.

## E.2 Task Description

Each task consists of:

- **Input:**
  - Query audio $A$
  - Paired audio ($A^p$ or $A^n$)
  - Generated short instruction (5–10 words)
- **Annotator Action:**
  1. Listen to both audios fully (replay if needed).
  2. Read the instruction carefully.
  3. Decide whether the instruction correctly captures the relationship between the audios.

## E.3 Annotation Labels

Annotators assign one of the following labels:

1. **Correct / Valid** – Instruction matches the difference or similarity.
   Example: $A$: "birdsong with distant car horn", $A^p$: "birdsong with louder car horn", Instruction: "louder car horn present".
2. **Incorrect / Invalid** – Instruction does not match the relation.
   Example: $A$: "guitar solo, clean tone", $A^p$: "guitar solo, distorted tone", Instruction: "softer background noise".
3. **Ambiguous / Unsure** – Instruction too vague, misleading, or unclear.
   Example: "different style".

## E.4 Special Consideration

- Instructions must remain within 5–10 words.
- For music: verify cues such as tempo, instrumentation, emotion, key, and production style.
- For environmental sounds: check loudness, background noise, reverb, and context.
- Cultural/era references (e.g., "90s rock style") must be plausible.

# F Examples

We list a few diverse examples from our Resonance set in Tab. 4.

# G Dataset Licenses

1. **AudioSet, AudioSet-Strong, and AudioCaps** - Released under the *Creative Commons Attribution 4.0 International (CC BY 4.0)* license for the metadata and feature embeddings. The associated YouTube audio clips are subject to YouTube's Terms of Service and are not freely redistributable.

```
I will provide you with captions and meta-data about two music samples: Music 1 (the
reference sample) and Music 2 (a semantically similar music but distinct sample). Your task
is to generate a clear, concise, retrieval instruction that captures the semantic
relationship between these two samples.
The generated instruction should enable retrieving Music 2 when given Music 1 and your
instruction as inputs. The instruction must highlight key differences or variations between
the two music samples in terms of mood, tempo, instrumentation, style, production quality,
or other musically relevant features.

Some examples of effective retrieval instructions include:
Tempo and Rhythm: "Something slower and with a deeper groove."
Genre-blending: "A track like this, but blending jazz with electronic elements."
Production Style: "Similar, but with acoustic guitar instead of electric."
Mood-based: "More upbeat."
Cultural or Regional Influence: "Similar vibe but influenced by Latin rhythms."
Era-specific: "With music like this but produced in a vintage, 80s style."
Structural or Arrangement-based: "Something with a stronger emphasis on the chorus."

These are just examples and feel free to come up with your own. Carefully analyze and
compare the two music samples from the data provided below to create your instruction. If
the two music samples are very similar or a good quality instruction is not possible, just
return "None". Quality should be the focus. **Do not always give too many hints about Music
1 in the instruction, when you address the similarity part. This means, you do not always
need to start with "Similar ...." -- only use this when necessary (when not using this does
not make sense to the final instruction or does not make it clear). You can directly start
with the final instruction like when deemed fit. Keep it as abstract as possible.** Here
are the two music samples:
Music 1: <audio>
Music 2: <audio>
```

Figure 3: Prompt used for instruction generation in the **music** domain. Given two music samples-Music 1 (reference) and Music 2 (semantically related but distinct)-the model is instructed to generate a short, abstract retrieval instruction that captures the key semantic difference between the two, focusing on attributes such as mood, tempo, instrumentation, or production style.

```
I will provide you with captions and meta-data about two audio samples: Audio 1 (the
reference sample) and Audio 2 (similar acoustic scene but distinct audio, possibly
containing unique acoustic events).
Your task is to generate a clear and concise retrieval instruction describing the
relationship between these samples based on their acoustic events or sound attributes.
The generated instruction should allow retrieving Audio 2 using Audio 1 and your
instruction as inputs, clearly highlighting key differences or variations.

Some examples of effective retrieval instructions include:
Intensity: louder bang
Mood: more lively atmosphere
Activity Level: fewer bird calls and calmer activity
Species Diversity: wider variety of bird species
Background Ambience: additional rustling leaves or water sounds
Environmental Context: similar soundscape but in an urban setting


These are just examples and feel free to come up with your own. Carefully analyze the data
for the 2 audio samples and compare the two of them provided below to create your
instruction. If the two audio samples are very similar or a good quality instruction is not
possible, just return "None". Quality should be the focus. **Do not always give too many
hints about Audio 1 in the instruction, when you address the similarity part. This means,
you do not always need to start with "Similar ...." -- only use this when necessary (when
not using this does not make sense to the final instruction or does not make it clear). You
can directly start with the final instruction like when deemed fit. Keep the instruction as
abstract as possible.**
Here are the two audio samples:
Audio 1: <audio>
Audio 2: <audio>
```

Figure 4: Prompt used for instruction generation in the **general sounds** domain. Given two audio samples-Audio 1 (reference) and Audio 2 (semantically related but distinct)-the model is instructed to produce a short, abstract retrieval instruction that highlights key differences in acoustic events or attributes (e.g., intensity, activity level, background ambience, environmental context) so that Audio 2 can be retrieved when paired with Audio 1.

| Music Samples | Sound Samples |
|---|---|
| More pronounced and repetitive metallic ringing. | More aggressive and distorted. |
| More vocal-driven with a funkier groove. | With a more defined lead vocal. |
| More rhythmic and percussive elements. | Missing the brief initial vocalization. |
| More spatial and less percussive. | More distorted and intense choral soundscape. |
| More sustained, orchestral horror atmosphere. | More driving, less vocally-sampled electronic music. |
| More upbeat with a funkier groove. | Additional sounds of commercial transactions. |
| Less driving and more atmospheric. | More vocal-led and deeply soulful. |
| Lacks musical instrumentation, primarily wind sounds. | More dynamic and contemporary band sound. |
| More percussive and deeper industrial texture. | Additional distinct mechanical clunks. |
| More grand and orchestral. | A cleaner, more melodically sung version. |
| More water sounds. | More minimalistic and hypnotic. |
| A faster, distorted electric rock version. | Less vocalization and more percussive sounds. |
| More expressive and ingredient-focused beverage review. | More acid-driven and hypnotic. |
| Adding sounds of cooking. | More atmospheric, with prominent electronic textures. |
| A slower, more subdued acoustic blues. | Smoother, more melodic classic hip-hop. |
| A more energetic, instrumental jazz piece. | More frenetic punk rock. |
| A broader, more cinematic jazz piece. | More distinct rhythmic elements. |
| Addition of clear human speech. | Slower, more resonant, and atmospheric. |
| More melodic pop-punk with female vocals. | More prominent high-pitched whine. |
| More aggressive singing, less rapping. | More epic and intense. |

Table 4: Diverse attributes from Resonance-Set.

2. **Clotho-v2** - Each audio file is sourced from Freesound under its respective Creative Commons license (primarily *CC BY* or *CC BY-NC*). The dataset metadata is distributed via Zenodo under the same open licenses.

3. **CochlScene** - Distributed under the *Creative Commons Attribution-ShareAlike 3.0 (CC BY-SA 3.0)* license according to its Zenodo record.

4. **MiraData** - Provided under the *GNU General Public License v3.0 (GPL-3.0)*. Redistribution and commercial use are restricted; the dataset is intended solely for academic and research purposes.

5. **Music4All** - Publicly available for research use. The extended Music4All-Onion subset is distributed under the *Creative Commons Attribution 4.0 (CC BY 4.0)* license.

6. **SoundDescs** – According to a dataset listing repository, SoundDescs is listed under "Apache, RemArc" licenses.

## H Use of AI assistants

We leveraged LLMs for three key aspects of our work: grammar and word choice refinement during the writing process, comprehensive literature searches to ensure proper citation of related work, and text data curation, consistent with common practices in LLM-related research.

## I Broader Impact

Resonance introduces the first framework for open-ended, instruction-based audio retrieval, enabling users to search for sounds and music based on free-form natural language descriptions of desired modifications or attributes. By unifying audio and textual representations through multimodal contrastive learning, Resonance opens new possibilities for semantic and intent-driven interaction with large audio collections.

The technology has broad applications across creative industries, accessibility tools, and intelligent media systems. For musicians and sound designers, Resonance can accelerate content discovery and remixing workflows by retrieving stylis-

tically related or complementary audio. In accessibility contexts, it can enhance audio browsing and summarization for individuals with hearing or cognitive impairments, aiding in assistive auditory technologies. Additionally, it contributes to the emerging ecosystem of large audio–language models, serving as a foundational resource for research in multimodal understanding, retrieval, and generation.

By releasing both the Resonance model and the Resonance-Set dataset, we aim to foster transparent and reproducible research in open-ended audio retrieval. The dataset's construction pipeline – comprising curated metadata, high-quality captions, and human-validated evaluation splits – can serve as a benchmark and community resource to drive further innovation in multimodal alignment and grounding.

**Potential Risk.** While Resonance presents clear benefits for creative and research applications, potential risks include the propagation of dataset biases across cultural or genre domains, inadvertent use of copyrighted material through similarity-based retrieval, and the environmental cost of large-scale model training. All audio data were sourced from public datasets under appropriate licenses, and our human evaluation process was conducted in compliance with institutional ethical standards.