ReasonAudio: A Benchmark for Evaluating Reasoning Beyond Matching in Text–Audio Retrieval

Anonymous ACL submission

Abstract

As multimodal content rapidly expands, audio retrieval has become essential for media search, content organization, and intelligent assistants. However, existing benchmarks mainly focus on semantic matching and overlook reasoning abilities required in real-world queries, such as negation, temporal ordering, concurrent events, and duration understanding. To address this gap, we introduce ReasonAudio, the first reasoning-intensive benchmark for Text–Audio Retrieval, comprising 1,000 queries and 10,000 composite audio clips across five fundamental reasoning tasks: Negation, Order, Overlap, Duration, and Mix. Despite their intuitive nature for humans and straightforward construction, these tasks pose significant challenges to current models. Our evaluation of ten state-of-the-art models reveals that all models struggle with reasoning-intensive audio retrieval, especially on Negation and Duration. Moreover, Multimodal Large Language Model-based embedding models fail to retain the reasoning capabilities of their backbones.[1]

1 Introduction

The rapid proliferation of multimodal content makes information retrieval increasingly challenging, highlighting the importance of multimodal retrieval in modern multimedia systems. As an important task, Text–Audio Retrieval aims to locate relevant audio clips from large-scale corpora based on textual queries, supporting applications such as media search (Elizalde et al., 2023), content organization (Liu et al., 2023), and intelligent assistants (Chen et al., 2025).

Existing Text–Audio retrieval methods can be broadly categorized into three paradigms. First, two-stage pipelines convert audio into text using captioning models (Huang et al., 2025; Chu et al., 2024) and then perform retrieval with text embedding models (Chen et al., 2024; Zhang et al., 2025c). While effective, their performance heavily depends on the quality of the intermediate textual representations. Second, direct embedding methods adopt CLIP-style contrastive learning (Radford et al., 2021) to align audio and text in a shared embedding space (Elizalde et al., 2023; Guzhov et al., 2022; Wu et al., 2022). Third, recent multimodal embedding approaches leverage pretrained MLLMs to learn unified cross-modal representations for retrieval (Xiao et al., 2025; Chen et al., 2026; Xu et al., 2025).

Despite these advances, most existing Text–Audio Retrieval benchmarks and methods focus primarily on semantic matching between audio and text, overlooking reasoning abilities required in real-world queries (SU et al., 2025; Zhang et al., 2025b,a; Alhamoud et al., 2025; Zhang et al., 2026), such as temporal reasoning, negation understanding, and contradiction detection. To address this limitation, we introduce ReasonAudio, the first Text–Audio Retrieval benchmark designed to evaluate logical reasoning beyond semantic matching. As shown in Figure 1, we collect 200 atomic sounds and synthesize 10,000 composite audio clips with controlled temporal patterns. Using predefined templates, we generate 1,000[2] reasoning-intensive queries and automatically annotate text–audio relevance. The benchmark evaluates four reasoning abilities: negation understanding, temporal ordering, concurrent event recognition, and duration discrimination.

Our comprehensive evaluation of 10 state-of-the-art (SOTA) models on ReasonAudio reveals that all methods struggle with reasoning-intensive retrieval tasks, with the best model, OmniEmbed-7B, achieving only 20.1% average accuracy. Further analysis reveals that MLLM-based models fail to effectively transfer pretrained reasoning capabili-

[1] Data and code will be released after review.

[2] Despite its modest size, 100 queries per task are widely accepted in information retrieval (Webber et al., 2008; Weller et al., 2025), and our pipeline is readily scalable.

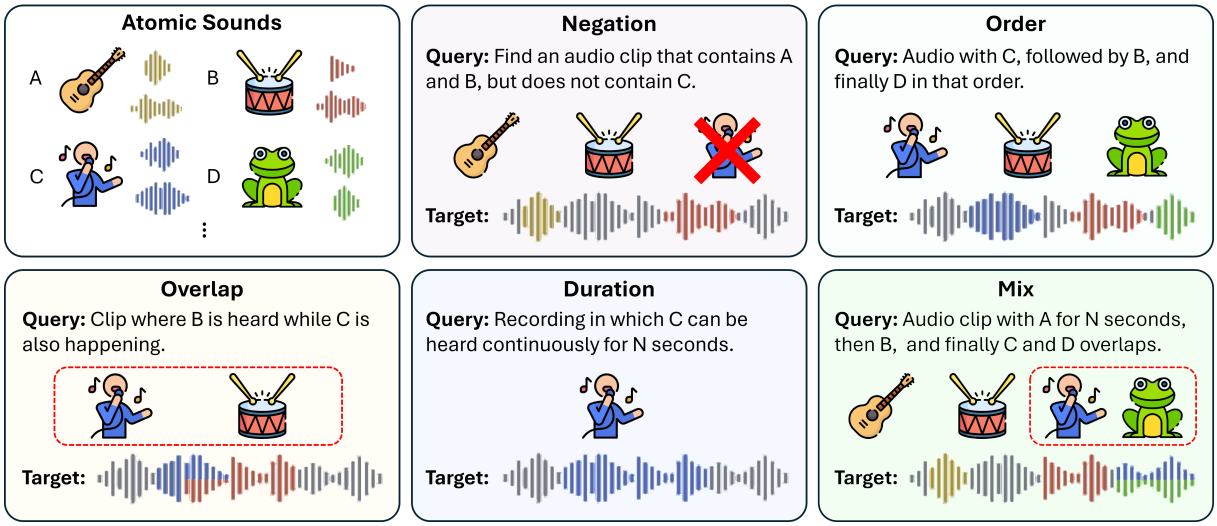

Figure 1: Illustration of the five reasoning-intensive cross-modal retrieval tasks in ReasonAudio. Each panel shows a sample query and a matching audio clip that satisfies its logical and temporal constraints.

ties and suffer from weak text-audio alignment in the shared embedding space. Our contributions are as follows:

- We introduce the first reasoning-intensive Text–Audio Retrieval benchmark designed to evaluate logical reasoning beyond semantic matching.
- We benchmark SOTA Text-Audio Retrieval models and show that current multimodal retrieval systems still struggle with reasoning-intensive retrieval tasks.

2 Task Formulation

We formally define the Text–Audio Retrieval task as follows. Let $Q = \{q_1, \ldots, q_n\}$ be text queries and $D = \{d_1, \ldots, d_m\}$ the audio corpus. For query $q_i \in Q$, each candidate $d_j \in D$ is either relevant (d_+) or non-relevant (d_-). The goal is to learn a similarity (scoring) function $f(q_i, d_j)$ such that, for a given query q_i, relevant audio candidate d_+ are ranked higher than irrelevant ones d_- in D. To evaluate audio reasoning beyond semantic matching, we introduce five tasks as shown in Figure 1:

- **Negation:** Retrieve audio clips containing target events while excluding undesired ones, requiring negation understanding.
- **Order:** Retrieve audio clips in which target events occur in a specified temporal sequence, requiring reasoning over temporal order.
- **Overlap:** Retrieve audio clips where target events temporally overlap, requiring reasoning about co-occurrence.
- **Duration:** Retrieve audio clips where an event persists for a specified duration, requiring reasoning about temporal extent.
- **Mix:** Retrieve audio clips that satisfy a combination of the above constraints, requiring multiple reasoning capabilities.

3 ReasonAudio Benchmark

We introduce ReasonAudio, a reasoning-intensive Text–Audio Retrieval benchmark built by synthesizing composite sounds and generating search queries from predefined templates. The dataset is constructed in three stages.

3.1 Atomic Sound Collection

To construct the audio corpus, we curated 200 atomic sound samples from FSD50K (Fonseca et al., 2022) and Freesound (Jiang et al., 2025). Each atomic sound corresponds to a single, clearly identifiable audio event. The collection process consists of four steps: (1) selecting high-quality candidate clips with clear and reliable labels; (2) manually reviewing annotations and standardizing filenames based on the provided labels; (3) extracting clean, continuous segments that fully capture the target sound event; and (4) normalizing all clips to a mono-channel WAV format. Importantly, we avoid semantically similar or hierarchical sound

Model	Negation		Order		Overlap		Duration		Mix		Avg.
	Acc@1	nDCG@10	Acc@1	nDCG@10	Acc@1	nDCG@10	Acc@1	nDCG@10	Acc@1	nDCG@10	
Two-Stage Text–Audio Retrieval											
Qwen2-Audio + BGE-M3	0.0	0.0	0.5	1.7	0.0	0.0	0.5	0.5	0.0	0.0	0.3
Qwen2-Audio + Qwen3-Embeding	0.0	0.0	0.5	0.5	0.0	0.0	1.5	0.7	0.0	0.0	0.3
Step-Audio + BGE-M3	2.0	1.4	0.5	0.5	2.5	3.8	8.5	7.4	1.0	2.4	3.0
Step-Audio + Qwen3-Embeding	4.0	2.7	1.5	2.4	4.0	5.8	6.0	7.5	3.0	4.9	4.2
CLIP-style Embedding Models											
CLAP	2.0	1.2	3.5	6.6	9.5	17.8	11.0	9.5	6.0	12.5	8.0
AudioCLIP	0.7	0.0	1.4	0.6	1.0	4.0	2.5	3.0	0.0	0.7	1.4
Wav2CLIP	0.0	0.2	0.0	0.0	0.0	0.0	1.0	1.8	0.0	0.2	0.3
MLLM-based Embedding Models											
LCO-Embedding-3B	3.0	5.0	16.0	26.2	21.5	29.5	10.5	10.5	6.0	14.8	14.3
LCO-Embedding-7B	2.0	2.8	23.0	31.8	**29.5**	**40.0**	**14.5**	**10.6**	15.5	25.1	19.5
e5-omni-7B	**10.0**	**10.0**	22.0	30.2	7.0	12.3	9.0	8.4	15.0	25.7	15.0
OmniEmbed-7B	5.5	7.0	**24.5**	**36.2**	18.0	26.4	6.5	8.5	**28.0**	**40.1**	**20.1**

Table 1: Comparison of SOTA Text–Audio Retrieval approaches and models on ReasonAudio. Performance is reported using Acc@1 and nDCG@10. The best and second-best results are shown in bold and underlined.

concepts (*e.g.,* "music" vs. "violin") to prevent ambiguity in query and relevance annotation.

3.2 Corpus and Query Synthesis

Based on atomic sounds, we synthesize both the audio corpus and search queries to construct reasoning-intensive retrieval tasks. We generate 10K composite audio clips by combining 2-8 atomic sounds under sequential and temporal-overlap compositions. Sequential compositions produce 6K clips with non-overlapping events and precise timestamps, while temporal-overlap compositions generate 4K clips with simultaneous events. Temporal relations and sound labels are recorded as metadata. Search queries are then generated using task-specific templates (Figure 1) by filling placeholders with metadata such as sound labels, temporal relations, durations, and negation constraints.

3.3 Relevance Annotation

We established ground-truth relevance using a deterministic program, followed by manual verification. The program labeled a query-audio pair as relevant only if the audio's composition metadata satisfied all constraints specified by the query. To ensure reliability, we manually reviewed 50 random queries for each task. This fully automated and scalable pipeline resulted in our final benchmark: ReasonAudio, which consists of 1K reasoning-intensive queries, 10K composite audio clips, and their relevance annotations.

4 Benchmarking SOTA Models

4.1 Experimental Setup

We evaluate three types of Text–Audio Retrieval approaches with ten SOTA models:

(1) **Two-Stage Text–Audio Retrieval.** We evaluated a two-stage approach that sequentially applies audio-to-text models and text retrievers. Audio inputs are first converted into text descriptions using **Qwen2-Audio** (Chu et al., 2024) or **Step-Audio** (Huang et al., 2025), followed by text retrieval with **BGE-M3** (Chen et al., 2024) or **Qwen3-Embedding** (Zhang et al., 2025c) based on the generated captions.

(2) **CLIP-style Embedding Models.** We evaluated CLIP-style multimodal embedding models trained with contrastive learning. We test **CLAP** (Elizalde et al., 2023), **AudioCLIP** (Guzhov et al., 2022), and **Wav2CLIP** (Wu et al., 2022), which differ in audio encoder architectures and pretraining data scales.

(3) **MLLM-based Embedding Models.** We evaluated unified multimodal retrieval embedding models, including **e5-omni-7B** (Chen et al., 2026), **LCO-Embedding-3B/7B** (Xiao et al., 2025), and **OmniEmbed-7B** (Xu et al., 2025).

Performance is evaluated using Accuracy@k and nDCG@k, with scores multiplied by 100.

4.2 Main Results

Reasoning-intensive tasks remain challenging for all evaluated models. As shown in Table 1, all models perform poorly on ReasonAudio, with

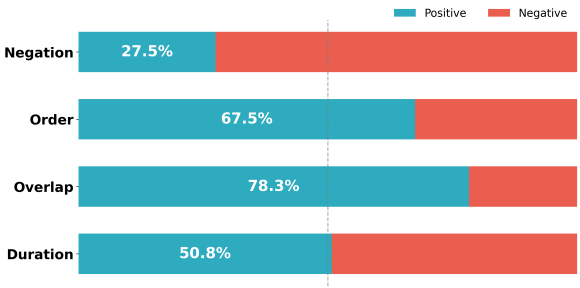

Figure 2: Performance of OmniEmbed-7B on the proposed multiple-choice retrieval task with two options.

best-performing OmniEmbed-7B achieving only 20.1% average accuracy. Two-stage and CLIP-style models are largely ineffective with Acc@1 below 10%. MLLM-based models show stronger performance—OmniEmbed-7B reaches 24.5% on Order and 28.0% on Mix, while LCO-Embedding-7B achieves 29.5% on Overlap—yet all struggle severely with Negation. These results indicate current models fail to handle sound matching under logical constraints, particularly for negation, temporal reasoning, and duration discrimination.

Two-stage approaches are ineffective due to audio captioning limitations. The two-stage pipeline performs worst, achieving only 0.3%–4.2% average accuracy. Replacing Qwen2-Audio with Step-Audio substantially improves performance, while stronger text retrievers bring minimal gains, indicating audio captioning is the main bottleneck. Manual inspection shows captions often fail to capture composite sound events accurately. Although Qwen2-Audio generates much longer captions than Step-Audio, verbosity does not improve retrieval; accurate core content is more critical than detailed but noisy descriptions.

MLLM-based models perform better but remain far from satisfactory. MLLM-based models achieve the strongest performance, with OmniEmbed-7B reaching 20.1% average accuracy, substantially outperforming other approaches. However, absolute performance remains low. While showing relatively stronger ability in temporal ordering and co-occurrence reasoning, they struggle with logical negation (2.0%–10.0%) and duration discrimination (6.5%–14.5%), suggesting they fail to inherit backbone reasoning capabilities. We hypothesize contrastive training emphasizes similarity matching over logical constraint satisfaction. Increasing model size yields moderate gains (LCO-Embedding-7B improves 5.2% over 3B), but no single model dominates across tasks, highlighting fundamental limitations in current multimodal embedding models for reasoning-intensive audio retrieval. Case studies are presented in Appendix C.

5 Analysis

5.1 How Well Do Models Reason When Audio Matching Is Factored Out?

ReasonAudio requires retrieval models to jointly perform sound matching and logical reasoning. To isolate reasoning ability, we further design a two-option retrieval setting where both candidates contain the queried sounds, but the negative candidate violates the logical constraint. For example, in a negation query such as "plastic bag and goose sounds, but without hailstorm," the negative candidate includes a hailstorm sound. We construct 120 questions per task. As shown in Figure 2, the best model, OmniEmbed-7B, performs poorly on Duration (50.8%) and Negation (27.5%), indicating weak logical reasoning and a bias toward sound matching, while achieving higher accuracy on Overlap (78.3%) and Order (67.5%).

5.2 How Well Are Text and Audio Embeddings Aligned in the Latent Space?

To qualitatively analyze text–audio alignment, we visualize five sampled queries using t-SNE in Appendix B. Positive audio samples fail to form compact clusters and are poorly separated from negatives. In several cases, negative samples lie closer to the query embedding than positives, leading to incorrect ranking. These results suggest that the model fails to encode logical constraints, especially negation, into the shared embedding space.

6 Conclusion

We introduce ReasonAudio, the first reasoning-intensive benchmark for Text–Audio Retrieval. Evaluating three retrieval paradigms and ten SOTA models reveals substantial limitations in current systems, especially on sound matching and audio reasoning tasks, where the best model achieves only 20.1% average accuracy. Our analysis further shows that MLLM-based embedding models fail to inherit the reasoning abilities of their backbones. Future work will extend the benchmark to more realistic settings.

Limitations

ReasonAudio relies on synthesized composite audio and template-generated queries, which may not fully reflect noisy real-world recordings, domain shift, or free-form user language. However, this design choice is primarily motivated by the need to ensure experimental controllability, scalability in benchmark construction, and accurate relevance annotation. Annotations follow deterministic rules from metadata; ambiguous or borderline cases in natural audio are not modeled here. The benchmark is built with English query templates; broader linguistic and cultural coverage is left to future work. We plan to develop a follow-up benchmark based on real-world music and speech audio and human-annotated queries incorporating logical constraints.

References

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip H.S. Torr, Yoon Kim, and Marzyeh Ghassemi. 2025. Vision-language models do not understand negation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29612–29622.

Haonan Chen, Sicheng Gao, Radu Timofte, Tetsuya Sakai, and Zhicheng Dou. 2026. e5-omni: Explicit cross-modal alignment for omni-modal embeddings.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, and Zhou Zhao. 2025. WavRAG: Audio-integrated retrieval augmented generation for spoken dialogue models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12505–12523, Vienna, Austria. Association for Computational Linguistics.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Soham Deshmukh, Benjamin Elizalde, and Huaming Wang. 2022. Audio retrieval with wavtext5k and clap training. *Preprint*, arXiv:2209.14275.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an audio captioning dataset. In *IEEE 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 736–740, United States. IEEE. IEEE International Conference on Acoustics, Speech and Signal Processing ; Conference date: 01-01-1900 Through 01-01-2000.

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.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. Fsd50k: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852.

Google Research. 2025. Simple voice questions (svq) dataset. https://huggingface.co/datasets/google/svq. Accessed: Feb. 1, 2026.

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.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, Peng Liu, Ruihang Miao, Wang You, Xi Chen, Xuerui Yang, Yechang Huang, Yuxiang Zhang, Zheng Gong, Zixin Zhang, Hongyu Zhou, Jianjian Sun, Brian Li, Chengting Feng, Changyi Wan, Hanpeng Hu, Jianchang Wu, Jiangjie Zhen, Ranchen Ming, Song Yuan, Xuelin Zhang, Yu Zhou, Bingxin Li, Buyun Ma, Hongyuan Wang, Kang An, Wei Ji, Wen Li, Xuan Wen, Xiangwen Kong, Yuankai Ma, Yuanwei Liang, Yun Mou, Bahtiyar Ahmidi, Bin Wang, Bo Li, Changxin Miao, Chen Xu, Chenrun Wang, Dapeng Shi, Deshan Sun, Dingyuan Hu, Dula Sai, Enle Liu, Guanzhe Huang, Gulin Yan, Heng Wang, Haonan Jia, Haoyang Zhang, Jiahao Gong, Junjing Guo, Jiashuai Liu, Jiahong Liu, Jie Feng, Jie Wu, Jiaoren Wu, Jie Yang, Jinguo Wang, Jingyang Zhang, Junzhe Lin, Kaixiang Li, Lei Xia, Li Zhou, Liang Zhao, Longlong Gu, Mei Chen, Menglin Wu, Ming Li, Mingxiao Li, Mingliang Li, Mingyao Liang, Na Wang, Nie Hao, Qiling Wu, Qinyuan Tan, Ran Sun, Shuai Shuai, Shaoliang Pang, Shiliang Yang, Shuli Gao, Shanshan Yuan, Siqi Liu, Shihong Deng, Shilei Jiang, Sitong Liu, Tiancheng Cao, Tianyu Wang, Wenjin Deng, Wuxun Xie, Weipeng Ming, Wenqing He, Wen Sun, Xin Han, Xin Huang, Xiaomin Deng, Xiaojia Liu, Xin Wu, Xu Zhao, Yanan Wei, Yanbo Yu, Yang Cao, Yangguang Li, Yangzhen Ma, Yanming Xu, Yaoyu Wang, Yaqiang Shi, Yilei Wang, Yizhuang Zhou, Yinmin Zhong, Yang Zhang, Yaoben Wei, Yu Luo, Yuanwei Lu, Yuhe Yin, Yuchu Luo, Yuanhao Ding, Yuting Yan, Yaqi Dai, Yuxiang Yang, Zhe Xie, Zheng Ge, Zheng Sun, Zhewei Huang, Zhichao Chang,

Zhisheng Guan, Zidong Yang, Zili Zhang, Binxing Jiao, Daxin Jiang, Heung-Yeung Shum, Jiansheng Chen, Jing Li, Shuchang Zhou, Xiangyu Zhang, Xinhao Zhang, and Yibo Zhu. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *Preprint*, arXiv:2502.11946.

Yuxuan Jiang, Zehua Chen, Zeqian Ju, Chang Li, Weibei Dou, and Jun Zhu. 2025. Freeaudio: Training-free timing planning for controllable long-form text-to-audio generation. *Proceedings of the 33rd ACM International Conference on Multimedia*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*.

A. Sophia Koepke, Andreea-Maria Oncescu, João F. Henriques, Zeynep Akata, and Samuel Albanie. 2023. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 25:2675–2685.

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. In *Interspeech*.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. AudioLDM: Text-to-audio generation with latent diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21450–21474. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Hongjin SU, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Liu Haisu, Quan Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O Arik, Danqi Chen, and Tao Yu. 2025. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. In *The Thirteenth International Conference on Learning Representations*.

Yulin Sun, Qisheng Xu, Yi Su, Qian Zhu, Yong Dou, Xinwang Liu, and Kele Xu. 2025. Audioset-r: A refined audioset with multi-stage llm label reannotation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 13089–13096. ACM.

William Webber, Alistair Moffat, and Justin Zobel. 2008. Statistical power in retrieval experimentation. In *Proceedings of the 17th ACM International Conference on Information and Knowledge Management*, pages 571–580. ACM.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2025. FollowIR: Evaluating and teaching information retrieval models to follow instructions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11926–11942, Albuquerque, New Mexico. Association for Computational Linguistics.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567.

Chenghao Xiao, Hou Pong Chan, Hao Zhang, Weiwen Xu, Mahani Aljunied, and Yu Rong. 2025. Scaling language-centric omnimodal representation learning. *Preprint*, arXiv:2510.11693.

Mengyao Xu, Wenfei Zhou, Yauhen Babakhin, Gabriel Moreira, Ronay Ak, Radek Osmulski, Bo Liu, Even Oldridge, and Benedikt Schifferer. 2025. Omni-embed-nemotron: A unified multimodal retrieval model for text, image, audio, and video. *Preprint*, arXiv:2510.03458.

Siyue Zhang, Yuan Gao, Xiao Zhou, Yilun Zhao, Tingyu Song, Arman Cohan, Anh Tuan Luu, and Chen Zhao. 2026. MRMR: A realistic and expert-level multidisciplinary benchmark for reasoning-intensive multimodal retrieval. In *The Fourteenth International Conference on Learning Representations*.

Siyue Zhang, Yuxiang Xue, Yiming Zhang, Xiaobao Wu, Anh Tuan Luu, and Chen Zhao. 2025a. MRAG: A modular retrieval framework for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3080–3118.

Siyue Zhang, Yilun Zhao, Liyuan Geng, Arman Cohan, Anh Tuan Luu, and Chen Zhao. 2025b. Diffusion vs. autoregressive language models: A text embedding perspective. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025c. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

A Dataset Comparison

Benchmark	Retrieval Type	Retrieval Task	Samples
AudioCaps (Kim et al., 2019)	Text ↔ Audio	Audio caption retrieval	4.9K
Clotho (Drossos et al., 2020)	Text ↔ Audio	Audio caption retrieval	5K
WavText5K (Deshmukh et al., 2022)	Text ↔ Audio	Audio caption retrieval	4.5K
SoundDescs (Koepke et al., 2023)	Text ↔ Audio	Audio caption retrieval	4.9K
AudioSet (Sun et al., 2025)	Audio → Text	Sound classification	0.5K
SVQ (MSEB) (Google Research, 2025)	Audio → Text	Voice search	181k
Spoken SQuAD (Lee et al., 2018)	Text → Audio	Spoken document retrieval	5.3K
ReasonAudio (Ours)	Text → Audio	Reasoning-intensive audio retrieval	1K

Table 2: Comparison of ReasonAudio with existing datasets for audio-related retrieval benchmarks.

B t-SNE Visualization

To qualitatively analyze the alignment between text and audio embeddings, we visualize five sampled queries using t-SNE in Figure 3. The results reveal substantial misalignment in the embedding space. Positive audio samples associated with the same query fail to form compact clusters, exhibiting highly dispersed distributions, while positive and negative samples are poorly separated. In several cases, negative samples lie closer to the query embedding than positive ones, leading to incorrect ranking. These observations indicate that the model fails to encode logical constraints—particularly negation—into the shared embedding space, resulting in insufficient discrimination between semantically valid and invalid audio candidates. This misalignment directly explains the below-random performance on negation tasks.

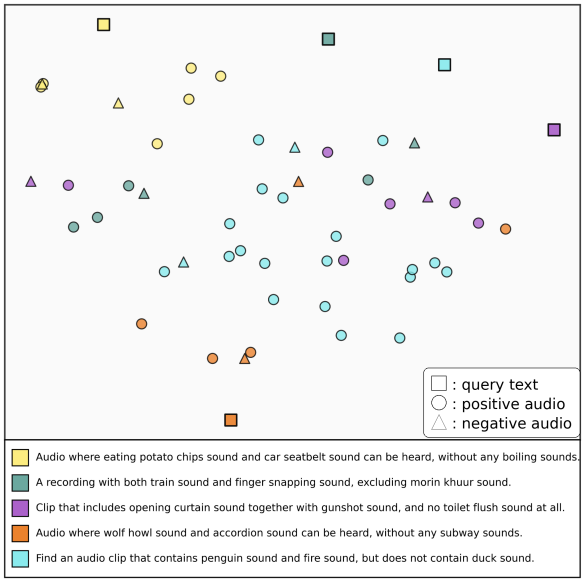

Figure 3: t-SNE plot of text and audio embeddings. Samples sharing the same color correspond to the same query.

C Case Studies

Tables 3–7 present additional OmniEmbed-7B top-5 retrieval cases for all five task types. Each table contains three queries and includes both successful and failed retrieval examples. For readability, each task-specific table only keeps the information relevant to that reasoning type.

Table 3: Negation cases. Each query requires the presence of both mentioned sounds; the table reports top-5 retrieved clips and relevance.

Case	Rank	Rel.	Retrieved audio
N1. Query: "Clip that has plastic bag sound and goose sound." Relevant audio in top-5: **No**.			
N1	1	No	Phone_Vibration_2s_Monkey_10s
N1	2	No	Mechanical_Fan_2s_Goose_2s
N1	3	No	Goose_5s_Tape_2s
N1	4	No	Goose_20s_Dishes_2s
N1	5	No	Penguin_2s_Cricket_2s
N2. Query: "Recording containing finger snapping sound and spray sound." Relevant audio in top-5: **Yes**, rank 1.			
N2	1	Yes	Spray_10s_Finger_Snapping_10s
N2	2	No	Cymbal_2s_Finger_Snapping_10s_Spray_15s
N2	3	No	Finger_Snapping_2s_Opening_Drink_2s
N2	4	No	Kiss_2s_Finger_Snapping_10s
N2	5	No	Drinking_Water_2s_Finger_Snapping_20s
N3. Query: "Find an audio clip that includes frog sound and ambulance sound." Relevant audio in top-5: **Yes**, rank 3.			
N3	1	No	Frog_15s_Scissors_5s
N3	2	No	Frog_15s_Maraca_5s
N3	3	Yes	Frog_20s_Ambulance_5s
N3	4	No	Frog_20s_Water_Droplet_10s
N3	5	No	Frog_15s_Sigh_15s

Table 4: Order cases. Each query specifies a temporal sequence; unrelated task-type details are omitted.

Case	Rank	Rel.	Retrieved audio
O1. Query: "Audio with harmonica sound, followed by engine sound, and finally sigh sound in that order." Relevant audio in top-5: **Yes**, rank 1.			
O1	1	Yes	Harmonica_15s_Engine_2s_Sigh_15s
O1	2	No	Harmonica_5s_Phone_Vibration_2s
O1	3	No	Harmonica_10s_Helicopter_2s
O1	4	No	Alarm_2s_Harmonica_10s_Yawn_15s
O1	5	No	Harmonica_20s_Cattle_5s_Dental_Drill_5s
O2. Query: "Recording that contains electric toothbrush sound, then cattle sound, and later erhu sound." Relevant audio in top-5: **No**.			
O2	1	No	Cattle_10s_Flute_15s
O2	2	No	Electric_Shaver_15s_Frog_10s
O2	3	No	Electric_Shaver_15s_Burping_20s
O2	4	No	Electric_Shaver_15s_Chewing_10s
O2	5	No	Electric_Guitar_10s_Chewing_5s
O3. Query: "Audio with screaming sound, followed by woodpecker sound, and finally knock sound in that order." Relevant audio in top-5: **Yes**, rank 2.			
O3	1	No	Woodpecker_2s_Screaming_10s
O3	2	Yes	Screaming_15s_Woodpecker_2s_Knock_5s
O3	3	No	Screaming_2s_Foot_Stamp_2s
O3	4	No	Owl_15s_Screaming_20s
O3	5	No	Seagull_5s_Screaming_10s_Cuckoo_5s

Table 5: Overlap cases. Each query asks for simultaneous sounds; temporal order, duration, and negation-specific fields are omitted.

Case	Rank	Rel.	Retrieved audio
V1. Query: "Audio with wind sound overlapping with aircraft sound." Relevant audio in top-5: **Yes**, rank 1.			
V1	1	Yes	Wind_10s_Aircraft_15s
V1	2	No	Aircraft_15s_Wind_5s
V1	3	No	Aircraft_15s_Drill_2s
V1	4	No	Aircraft_15s_Toilet_Flush_5s
V1	5	No	Fire_10s_Aircraft_15s
V2. Query: "Clip where turn page sound is heard while printer sound is also happening." Relevant audio in top-5: **No**.			
V2	1	No	Turn_Page_10s_Boiling_Water_5s
V2	2	No	Printer_10s_Finger_Snapping_2s
V2	3	No	Printer_10s_Gears_2s
V2	4	No	Balloon_Popping_5s_Printer_5s
V2	5	No	Printer_15s_Clock_Hands_2s
V3. Query: "Find an audio clip where drill and ship sound happen at the same time." Relevant audio in top-5: **Yes**, rank 2.			
V3	1	No	Truck_5s_Ship_15s_Foghorn_20s
V3	2	Yes	Drill_20s_Ship_20s
V3	3	No	Ship_15s_Woodpecker_2s
V3	4	No	Siren_10s_Truck_20s
V3	5	No	Electric_Toothbrush_5s_Seagull_5s_Foghorn_10s

Table 6: Duration cases. Each query specifies a target sound and duration; non-duration reasoning fields are omitted.

Case	Rank	Rel.	Retrieved audio
D1. Query: "Clip where the sound of drinking water sound lasts around 10 seconds." Relevant audio in top-5: **No**.			
D1	1	No	Filling_Liquid_10s_Filling_Liquid_5s
D1	2	No	Drinking_Water_5s_Gargling_Liquid_10s
D1	3	No	Crow_5s_Gargling_Liquid_5s
D1	4	No	Filling_Liquid_10s_Gargling_Liquid_10s
D1	5	No	Filling_Liquid_5s_Snake_2s
D2. Query: "Recording in which police car sound can be heard continuously for 20 seconds." Relevant audio in top-5: **Yes**, rank 1.			
D2	1	Yes	Siren_5s_Police_Car_20s
D2	2	No	Police_Car_10s_Maraca_10s
D2	3	No	Ambulance_10s_Motorcycle_2s
D2	4	No	Ambulance_20s_Crow_15s
D2	5	No	Police_Car_10s_Pipe_Organ_15s
D3. Query: "Recording in which fire sound can be heard continuously for 15 seconds." Relevant audio in top-5: **Yes**, ranks 3 and 5.			
D3	1	No	Fire_5s_Scissors_5s
D3	2	No	Clipping_Fingernail_5s_Fire_10s_Snoring_20s
D3	3	Yes	Prompt_Tone_2s_Fire_15s
D3	4	No	Fire_20s_Snake_2s
D3	5	Yes	Fire_15s_Wolf_Howl_2s

Table 7: Mix cases. Each query combines multiple reasoning requirements; task-specific fields from the single-reasoning tables are omitted to keep the examples compact.

Case	Rank	Rel.	Retrieved audio
M1. Query: "Find a clip where thunder and foghorn overlap, followed later by gasp and clock hands." Relevant audio in top-5: **Yes**, rank 1.			
M1	1	Yes	Thunder_15s_Foghorn_15s
M1	2	No	Bus_2s_Erasing_Board_2s_Thunder_10s
M1	3	No	Thunder_5s_Siren_5s
M1	4	No	Thunder_10s_Kiss_5s
M1	5	No	Thunder_15s_Spray_2s
M2. Query: "Find a clip where piano sound occurs before rain sound, and the final sniff sound lasts about 15 seconds." Relevant audio in top-5: **Yes**, rank 1.			
M2	1	Yes	Piano_2s_Rain_5s_Sniff_15s
M2	2	No	Musical_Keyboard_20s_Rain_10s
M2	3	No	Sniff_10s_Musical_Keyboard_15s
M2	4	No	Ukulele_20s_Sniff_15s
M2	5	No	Electric_Piano_2s_Rain_10s
M3. Query: "Find a clip where writing and subway overlap, while other retrieved sounds should not replace either event." Relevant audio in top-5: **Yes**, ranks 1 and 2.			
M3	1	Yes	Subway_10s_Writing_10s
M3	2	Yes	Writing_20s_Subway_10s
M3	3	No	Mallet_Percussion_2s_Writing_10s
M3	4	No	Writing_15s_Turn_Page_10s
M3	5	No	Writing_10s_Chewing_2s