WavRAG: Audio-Integrated Retrieval Augmented Generation for Spoken Dialogue Models

Yifu Chen [1,†] **Shengpeng Ji** [1,†] **Haoxiao Wang** [1,†] **Ziqing Wang** [3] **Siyu Chen** [1]
Jinzheng He [2] **Jin Xu** [2] **Zhou Zhao** [1,*]

[1] Zhejiang University [2] Alibaba Group [3] Beijing University of Technology
[†] Equal contribution. [*] Corresponding author.
Eve106298@163.com
zhaozhou@zju.edu.cn

Abstract

Retrieval Augmented Generation (RAG) has gained widespread adoption owing to its capacity to empower large language models (LLMs) to integrate external knowledge. However, existing RAG frameworks are primarily designed for text-based LLMs and rely on Automatic Speech Recognition to process speech input, which discards crucial audio information, risks transcription errors, and increases computational overhead. Therefore, we introduce WavRAG, the first retrieval augmented generation framework with native, end-to-end audio support. WavRAG offers two key features: 1) Bypassing ASR, WavRAG directly processes raw audio for both embedding and retrieval. 2) WavRAG integrates audio and text into a unified knowledge representation. Specifically, we propose the WavRetriever to facilitate the retrieval from a text-audio hybrid knowledge base, and further enhance the in-context capabilities of spoken dialogue models through the integration of chain-of-thought reasoning. In comparison to state-of-the-art ASR-Text RAG pipelines, WavRAG achieves comparable retrieval performance while delivering a 10x acceleration. Furthermore, WavRAG's unique text-audio hybrid retrieval capability extends the boundaries of RAG to the audio modality.

1 Introduction

Retrieval Augmented Generation (RAG) (Guu et al., 2020) has revolutionized natural language processing, offering a powerful approach to enhance text-based large language models (LLMs). A typical RAG system comprises three main components: a knowledge source (e.g., a database of documents or a knowledge graph) (Peng et al., 2024), a retriever module, and a generator module (typically a large language model). The RAG process unfolds in two primary stages. 1) Retrieval: Given an input query, the retriever module identifies and retrieves relevant information from the knowledge source.

2) Generation: The retrieved information, along with the original input query, is provided as context to the generator, which then produces the final response. This process allows the LLM to leverage up-to-date and factual knowledge, significantly improving the accuracy, consistency, and relevance of its responses (Gao et al., 2024). Researchers have sought to extend these benefits to the spoken dialogue domain (Lee et al., 2015; Chelba et al., 2008), aiming to enhance spoken dialogue models capable of processing audio inputs and generating speech responses (Fang et al., 2024; Ji et al., 2024a). However, these efforts have largely relied on cascaded "ASR+RAG" pipelines, which first transcribe speech to text using Automatic Speech Recognition (ASR) and then apply a text-based RAG system. This indirect approach suffers from several critical limitations: it fails to fully leverage the rich information present in the audio modality, treating it as a mere intermediary; the ASR component introduces computational overhead and potential transcription errors that propagate through the system; and the reliance on text-centric knowledge bases restricts the system's ability to utilize audio-specific knowledge. Crucially, the audio modality encompasses far more than just human speech; it includes a vast range of sounds (Ji et al., 2024b), such as environmental noises, music, and animal vocalizations, many of which are beyond the capabilities of ASR. A RAG framework that natively integrates this broader spectrum of audio information can unlock significant potential for richer, more contextually relevant understanding and generation (Chen et al., 2022), directly addressing key challenges of spoken dialogue models, such as the tendency to generate hallucinated content. However, realizing a fully end-to-end, audio-compatible RAG system remains a significant and open research challenge.

Addressing the limitations of existing approaches requires a fundamental shift: building knowledge bases that encompass a wide range of
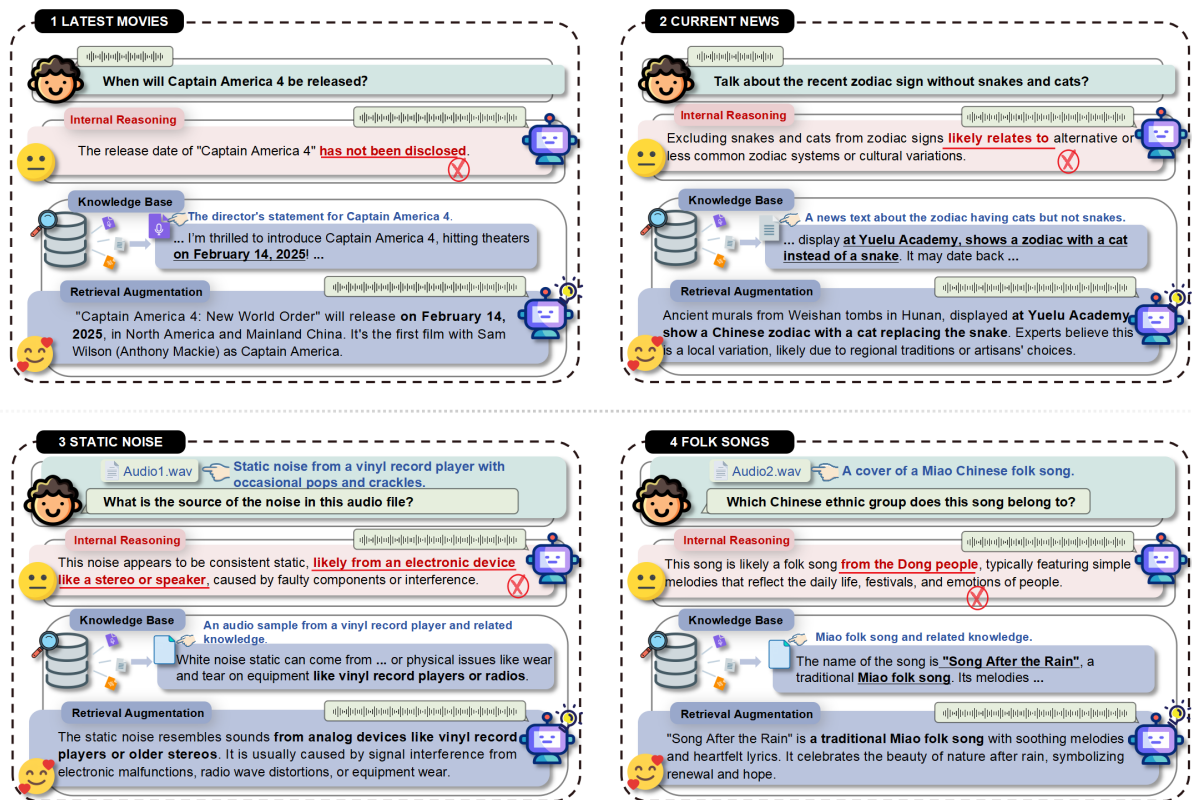
Figure 1: Dialogue examples illustrating WavRAG's ability to understand queries and generate appropriate responses by retrieving and augment relevant diverse modality knowledge

audio modalities alongside text, and developing retrieval mechanisms that can effectively represent and retrieve information from this unified audio-text space. Additionally,effectively harnessing retrieved multimodal information during generation requires new techniques for improved accuracy, naturalness, and contextual consistency. Therefore, we propose WavRAG, a novel end-to-end RAG framework designed for native audio integration. We show several dialogue scenarios to help understand the role of our framework in Figure 1. Inspired by LLM2Vec's (BehnamGhader et al., 2024) success in fine-tuning LLMs for text embeddings, we build our retriever on top of Qwen2-Audio, an MLLM with strong general audio comprehension, to create a unified embedding space for audio (speech and non-speech) and text. Considering that The pre-training objectives of multimodal language models are not optimized for retrieval, we further enhance the model with a contrastive learning framework. This approach allows the resulting retriever to encode end-to-end, directly encoding raw audio and text inputs into a shared embedding space, thus avoiding the computational overhead and potential error propagation of cascaded ASR-Text pipelines.

Furthermore, in the generation stage, We incorporate Chain-of-Thought (CoT) reasoning, promoting a structured and interpretable inference process that enhances both reliability and controllability in utilizing retrieved multimodal knowledge. In summary, our contributions are as follows:

- We propose WavRAG, a novel RAG framework for spoken dialogue models. It is the first to extend RAG to this domain in an end-to-end manner and to incorporate a hybrid text-audio knowledge base.

- We introduce a novel retriever WavRetriever, to support hybrid retrieval across text-audio modalities, and further enhance the in-context capabilities of the spoken dialogue models through Chain-of-Thought techniques.

- WavRAG achieves comparable results to the SOTA text-based RAG models in text retrieval, while offering an average acceleration of 10 times. Moreover, hybrid text-audio retrieval provides WavRAG with new capabilities.

2 Related Works

Audio RAG. While Retrieval-Augmented Generation (RAG) has shown promise in audio-related tasks like captioning (Koizumi et al., 2020; Zhao et al., 2023), text-to-audio generation (Huang et al., 2023), and music generation (Gonzales and Rudzicz, 2024). However, while these efforts demonstrate the utility of retrieval in audio processing, prior work primarily utilizes retrieval to enhance specific, isolated tasks with limited exploration of how retrieval-augmented techniques can benefit spoken dialogue models. Audio information itself carries rich semantic and acoustic imformation that can improve retrieval grounding, enhance response contextualization, and strengthen factual consistency. WavRAG, in contrast, integrates retrieval as a core component of a complete dialogue system. The combination of general audio support and end-to-end integration distinguishes WavRAG and represents a significant advancement towards truly audio-native, retrieval-augmented spoken dialogue systems.

Multimodal Retrieval. The increasing prevalence of multimedia applications and Retrieval-Augmented Generation (RAG) systems, fueled by Multimodal Language Models (MLLMs), has underscored the necessity for unified retrieval frameworks capable of managing diverse modalities. Traditional cross-modality retrieval methods often rely on pre-trained models such as CLAP (Elizalde et al., 2022) and CLIP, which use separate encoders for text and other modalities (e.g., UniVL-DR (Liu et al., 2023) and UniIR (Wei et al., 2023)). Other approaches enhance pre-trained text embeddings with audio encoders (Min et al., 2025), but these often prioritize the semantic content of speech, overlooking important general audio. Such methods struggle to effectively capture the full spectrum of information from both speech and non-speech audio. Recent advancements have highlighted the potential of Large Language Models (LLMs) and Supervised Fine-Tuning (SFT) for creating powerful, unified text representations (BehnamGhader et al., 2024; Lee et al., 2025). This methodology has been successfully extended to other modalities, with works like E5-V (Jiang et al., 2024) and VLM2VEC (Jiang et al., 2025) focusing on fine-tuning strategies for visual models. Furthermore, Zhang's research (Zhang et al., 2024) demonstrates the feasibility of developing universal multimodal retrieval models using MLLMs. However, there

has been limited exploration in the audio modality, prompting us to propose the first end-to-end audio-text multimodal retriever.

3 WavRAG

3.1 Overview

Figure 2 provides an overview of the traditional text-based RAG framework (top) and our proposed WavRAG framework (bottom).

In the context of text-based dialogue models, the classic retrieval-augmented generation (RAG) framework, depicted in the top portion of Figure 2, typically includes: (1) a text embedding model acting as the retriever R_ϕ, (2) a text-based dialogue model (Touvron et al., 2023) serving as the generator G_θ, and (3) a fixed external knowledge corpus $\mathcal{D} = \{d_1, \ldots, d_N\}$ containing only textual snippets d_i. During inference, the process is divided into two stages: retrieval and generation.

During retrieval, given a textual query q_t, the retriever computes the retrieval distribution $p(d \mid q_i)$ via:

$$p(d \mid q_i) = \frac{\exp(\text{sim}(R_\phi(q_i), R_\phi(d)))}{\sum_{d_i \in \mathcal{D}} \exp(\text{sim}(R_\phi(q_i), R_\phi(d_i)))}$$

where $R_\phi(\cdot)$ denotes the encoding function of R_ϕ, $\text{sim}(\cdot, \cdot)$ is a similarity metric (such as cosine similarity, as shown in Figure 2), and $d \in \mathcal{D}$. For efficiency, $\mathcal{D}$ is usually pre-encoded offline by R_ϕ. By drawing from this distribution, the system selects the Top-k relevant snippets, forming a subset $D_k \subset \mathcal{D}$.

During generation, conditioned on q_i and the retrieved subset D_k, the system computes the probability of producing the target text y_i as follows:

$$p(y_i \mid q_i, D_k) = \prod_{m=1}^{N} p(y_i \mid q_i, D_k, y_{<m})$$

where $p(y_i \mid q_t, D_k, y_{<i})$ is given by the generator G_θ and N denotes the number of tokens in the answer y_i. This process is illustrated in the top-right portion of Figure 2, where the retrieved text snippets and the query are used to prompt the LLM generator.

To extend RAG to spoken dialogue scenes and overcome the limitations of the traditional ASR-based approach, we propose WavRAG, as shown in the bottom portion of Figure 2. WavRAG integrates a retriever R_ϕ capable of directly processing queries in audio, text, and combined formats,
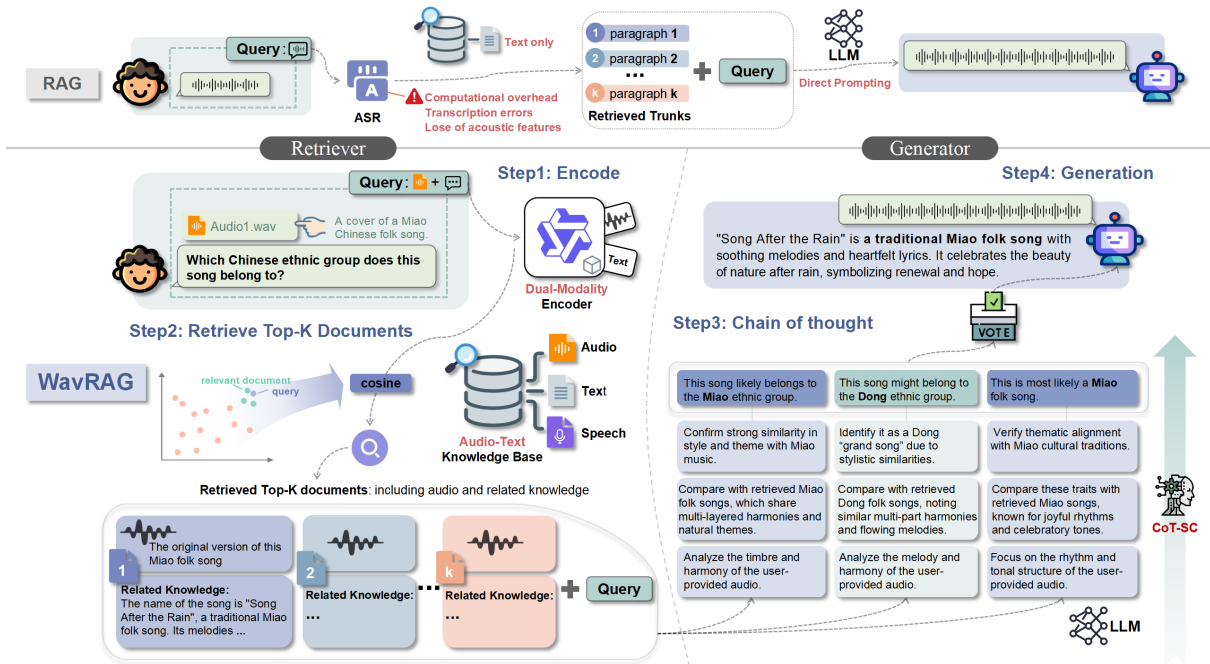
Figure 2: Architecture of the WavRAG framework. Top: Traditional RAG pipeline using ASR, highlighting its limitations. Bottom: WavRAG's four-step process: (1) A dual-modality encoder creates embeddings for both audio and text queries; (2) Top-K documents are retrieved from an audio-text knowledge base using cosine similarity; (3) A chain-of-thought reasoning process analyzes the retrieved information; (4) A large language model generates the final response, grounded in the retrieved knowledge.

and interfacing with a multimodal knowledge corpus K. This design preserves the full information present in the audio, including both speech and non-speech sounds. Specifically, we extend the original textual query q_t and the knowledge corpus $\mathcal{D}$ into a unified query q_{uni} and knowledge corpus $K = \{k_1, \ldots, k_i\}$, each of which may consist of audio, text, or a combination of these modalities, as depicted in the "Retriever" and "Knowledge Base" sections of Figure 2 (bottom). At the generation stage (bottom-right of Figure 2), we further introduce Chain-of-Thought (CoT) reasoning to systematically integrate external knowledge with the original input, ultimately producing the final answer. We will elaborate on the retrieval and generation components in detail in the following sections.

3.2 WavRetriever

The goal of our retriever, R_ϕ (WavRetriever), is to produce embedding vectors for both queries and knowledge entries that enable efficient similarity-based retrieval. As depicted in Figure 3, WavRetriever processes text, speech, or multimodal inputs, which are concatenated with a task-specific instruction and an End-of-Sequence (EOS) token. We build WavRetriever upon the Qwen2-Audio

MLLM, leveraging its robust general audio comprehension. Specifically, we freeze the pre-trained audio encoder parameters of Qwen2-Audio and focus training on the projection layer and the backbone LLM. This allows us to capitalize on Qwen2-Audio's existing audio processing capabilities.

However, simply fine-tuning Qwen2-Audio on the downstream task is insufficient for optimal retrieval performance. While pre-trained MLLMs like Qwen2-Audio possess robust multimodal understanding, their pre-training objectives are not directly optimized for creating embeddings suitable for similarity-based retrieval. To address this, we further adapt Qwen2-Audio into a powerful multimodal encoder using a carefully designed contrastive learning strategy. This strategy shapes the embedding space by maximizing the similarity between a query's embedding and its relevant (positive) knowledge, while minimizing the similarity with irrelevant (negative) knowledge embeddings.

Each training instance in our contrastive learning setup comprises a query, $q_{ins} = Instruction: \{prompt\}\ Query: q_{uni}$, a positive knowledge sample k^+, and a set of negative knowledge samples $\{k_1^-, \ldots, k_l^-\}$. Both the query, q_{uni}, and the knowledge samples, k,

can be audio, text, or a combination thereof. We construct negative samples using in-batch negatives. The representation for each instance is derived from the final hidden state of the last token. We employ the InfoNCE loss function (van den Oord et al., 2019), formulated as:

$$Z = \sum_{i=0}^{t} \exp\left(\frac{\text{sim}(r_q, r_{k,i})}{\tau}\right) \quad (1)$$

$$\mathcal{L} = -\left[\frac{\text{sim}(r_q, r_k^+)}{\tau} - \log Z\right] \quad (2)$$

where $sim(\cdot, \cdot)$ denotes cosine similarity, τ is a temperature parameter, and r_q represents the embedding of the query obtained after processing by the retriever R_ϕ. In Equation (1), the index $i = 0$ represents the embedding of the positive knowledge sample r_k^+, while the other i values correspond to the embeddings of the i-th negative knowledge sample $r_{k,i}^-$. The multimodal knowledge entries are stored in the knowledge base, as shown at the bottom of Figure 3. Detailed training setup and dataset-specific prompts are provided in Appendix A.
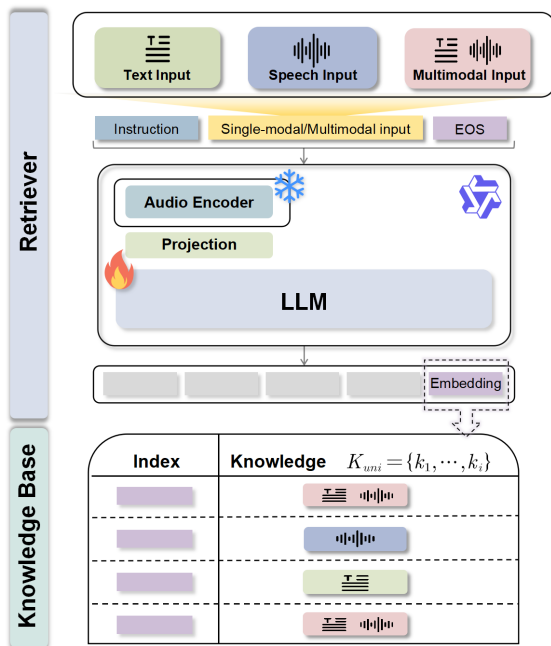

Figure 3: Architecture of the proposed multimodal retriever, showing the input processing, LLM-based encoding, and knowledge base structure.

3.3 Generation

In WavRAG's generation stage, we adopt a retrieval-augmented generation paradigm. The re-trieve provides the top-k retrieved knowledge entries (which can be audio, text, or multimodal) along with the original query q_{uni} as input to the generator. While this approach provides rich contextual information, most existing spoken dialogue systems are not trained on these lengthy and multiple mixed-modality input formats, making naive concatenation of all retrieved documents prone to suboptimal performance. To address this, we incorporate Chain-of-Thought (CoT) reasoning, specifically Zero-Shot-CoT (Kojima et al., 2022) and a Self-Consistency mechanism.

Zero-Shot-CoT Reasoning. Zero-Shot-CoT prompting leverages the in-context reasoning abilities of large language models (LLMs) to generate intermediate reasoning steps without requiring task-specific training examples. Given the multimodal query q_{uni}, a guiding prompt P_{prompt}, a "magic prompt" P' (e.g., "Let's think step-by-step"), and the top-k retrieved knowledge snippets $\mathcal{K}_k$, the generator $G_{\text{reasoning}}$ produces a reasoning chain C_{answer}:

$$C_{\text{answer}} = G_{\text{reasoning}}(q_{\text{uni}}, P_{\text{prompt}} + P', \mathcal{K}_k) \quad (3)$$

The retrieved knowledge $\mathcal{K}_k$ provides the context for the reasoning process, allowing the model to generate a logical, step-by-step deduction leading to the final answer.

Self-Consistency. To further enhance the reliability of the reasoning process, we employ a Self-Consistency mechanism. This approach samples multiple reasoning paths from the LLM and then selects the most consistent answer among them. This mitigates the risk of errors that can arise from relying on a single, potentially suboptimal, reasoning path. Specifically, we use the Universal Self-Consistency (USC) method (Chen et al., 2023). Instead of simply taking a majority vote among the generated answers (which can be problematic for free-form answers), USC concatenates all sampled reasoning paths and answers and prompts the LLM *itself* to select the most consistent response. This leverages the LLM's own understanding to determine the best answer, given the multiple reasoning paths.

12509

Task	Dataset	Model	Whisper size	WER	Avg. Time	Metric			
						R@1	R@5	R@10	NDCG@10
Speech2Text	HotpotQA	BGE	Tiny	37.55%	1.26	0.3741	0.7024	0.7509	0.4628
			Medium	21.67%	1.48	0.4440	0.8319	0.8736	0.5190
			Large	19.2%	1.92	**0.4533**	**0.8519**	0.8895	**0.5252**
		Ours	-	-	**0.23**	0.4532	0.8492	**0.8898**	0.5117
Comparison vs. BGE (Tiny / Medium / Large):		*Speed-up* ≈ 5.49 × /6.43 × /8.35×,				∆R@10 ≈ +0.139/ + 0.016/ + 0.0003.			
Text2Speech	Spoken-SQuAD	CLSR	-	-	-	0.4982	0.7963	0.8583	-
		BGE	-	44.22%	-	0.5464	0.7767	0.8497	0.6947
		Ours	-	-	**0.11**	**0.6844**	**0.8374**	**0.9023**	**0.8483**
Speech2Speech	SLUE-SQA-5	CLSR	-	16.69%		0.3065	0.6219	**0.7443**	-
		BGE	Tiny	45.34%/53.66%	0.62/1.27	0.1696	0.3871	0.4828	0.2194
			Medium	26.14%/44.46%	0.87/3.44	0.3228	0.5940	0.6982	0.2989
			Large	23.59%/42.19%	0.98/4.63	0.3312	0.6121	0.7196	0.3269
		Ours	-	-	**0.17/0.22**	0.3392	0.6308	0.7221	**0.3623**
Comparison vs. BGE (Tiny / Medium / Large):		*Speed-up* ≈ 4.84 × /11.05 × /14.38×,				∆R@10 ≈ +0.2393/ + 0.0282/ + 0.0025.			
Audio+Text2Audio+Text	Ours	ClAP (AT)	-	-	**0.05**	0.1260	0.2940	0.3989	0.2474
		CLAP(TA)	-	-	**0.05**	0.0998	0.2577	0.3588	0.2135
		CLAP(AT2AT)	-	-	0.09	0.1345	0.2145	0.2379	0.1849
		ClAP (ALL)	-	-	0.06	0.0001	0.0012	0.0018	0.0002
		BGE (Caption)	-	-	1.99	0.0251	0.0585	0.0775	0.0483
		Ours	-	-	0.19	**0.2728**	**0.5184**	**0.6313**	**0.4381**

Table 1: Comparison of various models and configurations across multiple tasks including Speech2Text, Text2Speech, Speech2Speech, and Audio+Text2Audio+Text . Performance metrics include Word Error Rate (WER), Average Time, and multiple retrieval metrics (R@1, R@5, R@10, and nDCG@10). CLAP are evaluated for their respective tasks: CIAP (AT) is tested only on the Audio-to-Text retrieval task subset, CLAP (TA) on the Text-to-Audio retrieval task, CLAP (AT2AT) on Audio-to-Text-to-Audio tasks, and CLAP (ALL) on the entire dataset. Speed-up and performance changes relative to BGE (different Whisper model size Tiny/Medium/Large) configurations are reported.

4 Experiments

4.1 Datasets

For training, we curated a dataset of 1.5M samples across five retrieval scenarios: **Speech-to-Text:** We adapted existing text retrieval datasets (e.g. HotpotQA (Yang et al., 2018), Quora (Wang et al., 2017)) by synthesizing speech queries using the CosyVoice2 TTS model (Du et al., 2024; An et al., 2024) with diverse voice prompts and noise augmentation. **Speech-to-Speech and Text-to-Speech:** We used existing datasets: SLUE-SQA-5 (Shon et al., 2023) and Spoken-SQuAD (Li et al., 2018). **Text-to-Text:** We used existing text retrieval datasets: ELI5 (Fan et al., 2019), NQ (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), MS MARCO, Quora, SQuAD (Rajpurkar et al., 2016), and TriviaQA (Joshi et al., 2017). **Audio+Text-to-Audio+Text:** We process new data from sources like AudioSetSL, Audio-Caps (Kim et al., 2019), MusicCaps (Agostinelli et al., 2023), Clotho (Drossos et al., 2019), Vox-Celeb (Nagrani et al., 2017), and Xeno-canto, where queries and documents are both general audio-text pairs. For evaluation, We test WavRAG

on four datasets: HotpotQA, Spoken-SQuAD, SLUE-SQA-5, and our custom mixed modality dataset. Detailed data processing procedures, dataset statistics, and examples are provided in Appendix B

4.2 Baselines

Retrieval Baselines. 1) **BGE** (Li et al., 2024): A state-of-the-art text embedding model, used within an ASR-based pipeline for speech-related retrieval tasks. 2) **CLSR** (Anonymous, 2025): A state-of-the-art speech-text retrieval framework. Used for comparison on speech-to-text and text-tp-speech retrieval task. 3) **CLAP** (Elizalde et al., 2022): Used for comparison on our custom multimodal dataset. 4) **Qwen2Audio-enhanced Text Retrieval**: Used on our custom dataset. This baseline leverages Qwen2Audio to generate descriptive text from audio clips, which is then concatenated with the original text by using the BGE model.

Generation Baselines. 1) **TextRAG**: A standard text-based RAG pipeline using BGE embeddings for retrieval and Whisper medium for ASR.

Method	Model	Input	EM		Avg EM	FS (Ours)
			HotpotQA	SLUE-SQA-5		
(a) TextRAG						
GPT-4o		top-1	0.3124	0.3237	0.3181	-
		top-2	0.3457	0.3359	0.3408	-
		top-3	0.3623	0.3531	0.3577	-
		Oracle	0.5853	0.5931	0.5892	-
QwenAudio		top-1	0.1783	0.2439	0.2111	-
		top-2	0.2336	0.2502	0.2419	-
		top-3	0.2417	0.2561	0.2489	-
		Oracle	0.4867	0.4784	0.4824	-
(b) WavRAG						
GPT-4o		top-1	0.4019	0.3904	**0.3962**	**0.5732**
		top-2	0.4186	0.4315	**0.4249**	**0.6408**
		top-3	0.4271	0.4007	**0.4139**	**0.5129**
		Oracle	0.5941	0.6164	**0.6053**	**0.7096**
QwenAudio		top-1	0.2033	0.2647	**0.2340**	**0.5387**
		top-2	0.2439	0.2956	**0.2698**	**0.5521**
		top-3	0.2658	0.3063	**0.2860**	**0.5387**
		Oracle	0.5032	0.5294	**0.5163**	**0.6079**
(c) WavRAG-CoT						
GPT-4o		top-1	0.4261	0.4520	**0.4390**	**0.6412**
		top-2	0.4286	0.5239	**0.4983**	**0.6487**
		top-3	0.4403	0.4918	**0.4662**	**0.5981**
		Oracle	0.5976	0.6849	**0.6413**	**0.7389**
QwenAudio		top-1	0.2688	0.3132	**0.2910**	**0.6386**
		top-2	0.3026	0.3352	**0.3189**	**0.6017**
		top-3	0.3152	0.3397	**0.3275**	**0.5612**
		Oracle	0.5863	0.6103	**0.5983**	**0.7122**

Table 2: Generation experiment results: Performance comparison of TextRAG, WavRAG, and WavRAG-CoT on HotpotQA and SLUE-SQA-5 datasets. Metrics include Exact Match (EM) and F1-Score (FS). Results are shown for both GPT-4o and QwenAudio base models, with varying numbers of retrieved documents (top-1, top-2, top-3, and Oracle).

4.3 Evaluation Metrics and Experimental Settings

Retrieval. Retrieval performance is evaluated across four scenarios: Speech-to-Text (HotpotQA), Speech-to-Speech (SLUE-SQA-5), Text-to-Speech (Spoken-SQuAD), and Audio+Text to Audio+Text (a custom dataset). Reported metrics include:

Recall@k: The proportion of relevant items found within the top-k retrieved results. Higher is better. (We report Recall@1, Recall@5, and Recall@10). NDCG@10 (Normalized Discounted Cumulative Gain): A measure of ranking quality that considers the position of relevant items in the retrieved list, giving higher scores to relevant items ranked higher. Higher is better. Average Inference Time: The average time taken to process a single query.

For scenarios involving speech input, the Word Error Rate (WER) and model size of the Whisper ASR model (used in baseline methods) are also reported. WER measures the accuracy of the speech recognition, with lower values indicating fewer errors.

Generation. Three RAG frameworks are compared: TextRAG (which uses Whisper medium for ASR and BGE embeddings for retrieval), WavRAG (using WavRetriever), and WavRAG-CoT (WavRAG with Chain-of-Thought reasoning). Both GPT-4o and QwenAudio serve as the generation models. The following metrics are used:

EM (Exact Match): For short-form answers (in HotpotQA and SLUE-SQA-5), this is a binary metric. It is 1 if the generated answer exactly matches the ground truth answer and 0 otherwise. Higher is better. FactScore: For long-form answers (in the custom dataset), this metric, utilizing Qwen-plus for fact verification, evaluates the factual accuracy of the generated text by assessing the proportion of factual claims in the generation that are supported by the retrieved evidence. Higher is better.

Results are presented for the top-1, top-2, and top-3 retrieved documents. This means the generation model is provided with the single most relevant retrieved document (top-1), the two most relevant documents (top-2), and the three most relevant documents (top-3), respectively. Results are also presented for an "Oracle" condition where only the ground truth document is used.

Dataset	Metric	Qwen2audio (Original)	WavRAG	Improvement
Ours	R@1	0.0675	**0.2728**	*+0.2053*
	R@5	0.1457	**0.5184**	*+0.3727*
	R@10	0.1868	**0.6313**	*+0.4445*
	nDCG@10	0.1212	**0.5381**	*+0.4169*
Spoken-SQuAD	R@1	0.3407	**0.6844**	*+0.3437*
	R@5	0.4995	**0.8374**	*+0.3379*
	R@10	0.6003	**0.9023**	*+0.3020*
	nDCG@10	0.3554	**0.8483**	*+0.4929*
HotpotQA	R@1	0.1457	**0.4532**	*+0.3075*
	R@5	0.3172	**0.8492**	*+0.5320*
	R@10	0.3858	**0.8898**	*+0.5040*
	nDCG@10	0.2868	**0.5117**	*+0.2249*

Table 3: Performance improvement after contrastive learing

4.4 Main Results

Retrieval Performance. Table 1 presents the retrieval results, demonstrating WavRAG's superior performance and efficiency compared to traditional ASR-dependent baselines and other approaches across all four evaluated tasks: Speech-to-Text, Text-to-Speech, Speech-to-Speech, and Audio+Text-to-Audio+Text.

WavRAG's key advantage lies in its direct processing of audio inputs, eliminating the need for ASR and its associated computational overhead and potential for transcription errors. This translates to significant speedups in inference time,
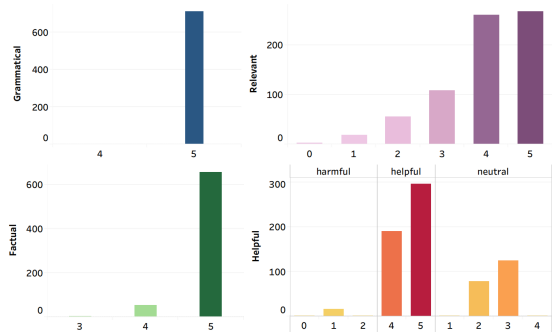
Figure 4: Human evaluation of knowledge quality. Distributions are shown for Grammatical scores, Factual scores, Relevance scores, and Helpfulness scores. The Helpfulness plot is further broken down by helpfulness level (helpful, neutral, harmful).

ranging from approximately 5x to over 14x compared to BGE models using Whisper, while simultaneously achieving comparable or superior retrieval accuracy. The most substantial performance gains are seen in the challenging Audio+Text-to-Audio+Text scenario, where WavRAG dramatically outperforms all baselines. This highlights the effectiveness of WavRAG's unified multimodal embedding space in capturing the complex relationships between audio and text, a capability lacking in approaches that rely on separate encoders or text-based representations. Even in scenarios where strong text-based baselines exist (e.g., Speech-to-Text), WavRAG achieves competitive results without the ASR bottleneck.

Generation Performance. Table 2 presents the generation experiment results. The results clearly demonstrate two key findings: 1) Impact of Direct Audio Input (WavRAG vs. TextRAG): Across all datasets and both LLMs, WavRAG consistently outperforms the traditional TextRAG approach. This highlights the benefits of providing the generator with direct access to the original audio modality, rather than relying on potentially lossy ASR transcriptions. For example, with GPT-4o on HotpotQA, WavRAG achieves a top-1 EM of 0.4019, compared to TextRAG's 0.3124 (an absolute improvement of +0.0895). Similar gains are observed on SLUE-SQA-5 and with QwenAudio. On our custom dataset, where FactScore is used, WavRAG (with GPT-4o) achieves a top-1 score of 0.5732, demonstrating a substantial availability for the inclusion of audio information. 2) Effectiveness of Chain-of-Thought (WavRAG-CoT): The addition of Chain-of-Thought (CoT) reasoning further en-

hances performance. WavRAG-CoT consistently outperforms standard WavRAG across all datasets and models. With GPT-4o, the top-1 EM increases from 0.3904 to 0.4520 on SLUE-SQA-5 (+0.0616). On our custom dataset, the FactScore improves from 0.5732 to 0.6412 (+0.068). A noteworthy observation is the performance trend across Top-k configurations. We observe a decrease in performance when moving from Top-2 to Top-3 in our dataset.This degradation suggests Models struggle to prioritize and synthesize information from a larger, more heterogeneous set of inputs.The fact that WavRAG-CoT mitigates this issue is significant. The structured, step-by-step reasoning enforced by CoT provide a mechanism for the model to better manage the complexity of multimodal knowledge.

4.5 Analysis

Ablation Studies on Contrasitive Training Framework. To isolate the impact of our contrastive learning framework, we conducted an ablation study comparing the fine-tuned WavRAG retriever to the pre-fine-tuned Qwen2-Audio-7B-Instruct model. This baseline represents a strong, pre-trained MLLM with inherent multimodal understanding, but without retrieval-specific optimization. Following (Jiang et al., 2023), original Qwen2-Audio's representation was obtained by prompting for global semantics at the last token.

Table 3 shows the results, with WavRAG significantly outperforming the baseline across all datasets and metrics. Recall@1 improvements range from +0.3075 to +0.3437, and nDCG@10 gains are even more pronounced, reaching up to +0.4169. These substantial improvements unequivocally validate the effectiveness of our contrastive learning framework in adapting the MLLM for multimodal retrieval.

Knowledge Extension Quality. A critical aspect of WavRAG is its ability to take advantage of extended knowledge associated with audio. To assess the quality of this generated knowledge, we conducted a human evaluation on 700 randomly sampled instances from our custom multimodal dataset. Fluent English-speaking annotators evaluated each sample on a 5-point scale across four key dimensions: Grammaticality , Factual accuracy, Relevance , and overall Helpfulness (further categorized as helpful, neutral, or harmful). Figure 4 shows the score distributions. The vast majority of

samples received a score of 5 for Grammaticality, Factual accuracy, and Relevance, indicating high-quality knowledge generation. Furthermore, most samples were rated as "Helpful," demonstrating the positive impact of the extended knowledge on audio understanding.

5 Conclusions

This work introduced WavRAG, a novel retrieval-augmented generation framework specifically designed for spoken dialogue systems. WavRAG makes a significant departure from traditional ASR-dependent pipelines by directly processing raw audio input for embedding and retrieval. This approach offers several key advantages, including reduced computational overhead, preservation of rich acoustic information, and the ability to leverage a unified multimodal knowledge base.Through comprehensive experiments, including quantitative evaluations and qualitative analyses , we demonstrated the effectiveness of WavRAG. The results show substantial improvements in both retrieval and generation performance compared to traditional methods and baseline models.

Limitations

Despite WavRAG's exploration of how a well-designed RAG system can leverage both semantic and acoustic information to enhance the semantic quality of responses, emotional tone and prosody are equally crucial in spoken dialogue systems. The extent to which RAG can contribute to the acoustic aspects of responses, such as intonation, expressiveness, and speaker style, remains an open question, warranting further investigation.

Acknowledgements

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. Musiclm: Generating music from text. *Preprint*, arXiv:2301.11325.

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, et al. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Anonymous. 2025. CLSR: End-to-end contrastive language-speech retriever for better speech retrieval augmented generation. In *Submitted to ACL Rolling Review - December 2024*. Under review.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *Preprint*, arXiv:2404.05961.

Ciprian Chelba, Timothy J. Hazen, and Murat Saraclar. 2008. Retrieval and browsing of spoken content. *IEEE Signal Processing Magazine*, 25(3):39–49.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *Preprint*, arXiv:2210.02928.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2019. Clotho: An audio captioning dataset. *Preprint*, arXiv:1910.09387.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *Preprint*, arXiv:2412.10117.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2022. Clap: Learning audio concepts from natural language supervision. *Preprint*, arXiv:2206.04769.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *Preprint*, arXiv:2409.06666.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Robie Gonzales and Frank Rudzicz. 2024. A retrieval augmented approach for text-to-music generation. In *Proceedings of the 3rd Workshop on NLP for Music and Audio (NLP4MusA)*, pages 31–36, Oakland, USA. Association for Computational Lingustics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *Preprint*, arXiv:2002.08909.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *Preprint*, arXiv:2301.12661.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Yidi Jiang, Jingzhen He, Yunfei Chu, Jin Xu, and Zhou Zhao. 2024a. Wavchat: A survey of spoken dialogue models. *Preprint*, arXiv:2411.13577.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al. 2024b. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2024c. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10301–10305. IEEE.

Shengpeng Ji, Jialong Zuo, Wen Wang, Minghui Fang, Siqi Zheng, Qian Chen, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, et al. 2024d. Controlspeech: Towards simultaneous zero-shot speaker cloning and zero-shot language style control with decoupled codec. *arXiv preprint arXiv:2406.01205*.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2023. Scaling sentence embeddings with large language models. *Preprint*, arXiv:2307.16645.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024. E5-v: Universal embeddings with multimodal large language models. *Preprint*, arXiv:2407.12580.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2025. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *Preprint*, arXiv:2410.05160.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuma Koizumi, Yasunori Ohishi, Daisuke Niizumi, Daiki Takeuchi, and Masahiro Yasuda. 2020. Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval. *Preprint*, arXiv:2012.07331.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Nv-embed: Improved techniques for training llms as generalist embedding models. *Preprint*, arXiv:2405.17428.

Lin-shan Lee, James Glass, Hung-yi Lee, and Chun-an Chan. 2015. Spoken content retrieval—beyond cascading speech recognition with text retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9):1389–1420.

Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024. Making text embedders few-shot learners. *Preprint*, arXiv:2409.15700.

Chia-Hsuan Li, 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. *Preprint*, arXiv:1804.00320.

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2023. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. *Preprint*, arXiv:2209.00179.

Do June Min, Karel Mundnich, Andy Lapastora, Erfan Soltanmohammadi, Srikanth Ronanki, and Kyu Han. 2025. Speech retrieval-augmented generation without automatic speech recognition. *Preprint*, arXiv:2412.16500.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. In *Interspeech 2017*, interspeech_2017. ISCA.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey. *Preprint*, arXiv:2408.08921.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan Sharma, Wei-Lun Wu, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. 2023. SLUE phase-2: A benchmark suite of diverse spoken language understanding tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8906–8937, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. *Preprint*, arXiv:1702.03814.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2023. Uniir: Training and benchmarking universal multimodal information retrievers. *Preprint*, arXiv:2311.17136.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. Gme: Improving universal multimodal retrieval by multimodal llms. *Preprint*, arXiv:2412.16855.

Jinming Zhao, Gholamreza Haffari, and Ehsan Shareghi. 2023. Generating synthetic speech from SpokenVocab for speech translation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1975–1981, Dubrovnik, Croatia. Association for Computational Linguistics.

A Training Detail and Task-specific Instructions

In our training regimen, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2021) with a rank of 8, a learning rate of 1e-4, an alpha of 32, and a dropout rate of 0.05. To efficiently handle audio segments of varying lengths while maintaining training efficacy, each audio sample is tokenized with a maximum of 2000 tokens, and the corresponding text is limited to 512 tokens, with a batch size of 64. Additionally, gradient checkpointing is utilized and training is conducted in bfloat16 precision to optimize GPU memory usage; all experiments are executed on four NVIDIA A800 GPUs, each equipped with 80GB of memory. To further enhance model robustness to diverse acoustic environments, we integrate several speech data augmentation techniques that sequentially apply echo simulation, MUSAN noise addition, and random gain adjustments. Specifically, the echo effect is simulated by adding a delayed (100–500 ms) and scaled (0–0.2) version of the audio signal to itself; MUSAN noise is incorporated by randomly selecting a noise file, resampling it to the target rate if necessary, concatenating it to match the audio length, and mixing it at a target signal-to-noise ratio between -4 dB and 14 dB with a probability of 0.5; finally, random gain is applied with a probability of 0.5, scaling the audio by a factor corresponding to a gain between -4 dB and 15 dB.

B Data process pipeline and examples

B.1 Date process

To address the challenges of audio-text retrieval across varied applications, we trained our model

Task	Dataset	Train Q-D pairs	Retrieval Test	Generation Test
Speech-to-Text	Quora	60202	-	-
	HotpotQA	84516	7405	7405
Text-to-Text	ELI5	325475	-	-
	TrivialQA	60315	-	-
	SQuAD	87599	-	-
	MS MARCO	485823	-	-
Speech-to-Specch	SLUE-SQA-5	46186	2382	2382
Text-to-Speech	SpokenSQuAD	37111	5351	
Audio+Text-to-Audio+Text	Ours	78746	8834	1200
	AudioCaps	35327	4043	572
	MusicCap	4080	442	76
(Ours)	Clotho	2852	314	68
	VoxCeleb	1091	120	-
	Xeno-canto	8771	956	-
	Collected	-	-	43
	Total	130867	14709	1959

Table 4: Training data

on four distinct retrieval scenarios: speech-to-text, speech-to-speech, text-to-speech and audio+text-to-audio+text,totally 1.5M samples, as shown in table 4. For each scenario, we meticulously designed a dedicated data construction pipeline, ensuring the training data appropriately reflects the task-specific nuances.

Speech-to-Text Retrieval   For the speech-to-text retrieval task, we leveraged a suite of established text retrieval datasets, including ELI5 (Fan et al., 2019), NQ (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), MS MARCO, Quora (Wang et al., 2017), SQuAD (Rajpurkar et al., 2016), and TriviaQA (Joshi et al., 2017). To transform these text-based queries into speech, we employed the CosyVoice2 (Du et al., 2024) text-to-speech model (Ji et al., 2024c,d). To further enrich the diversity and robustness of the synthesized speech, we incorporated several data augmentation techniques. Specifically, we randomly sampled voice prompts from the Common Voice 12.0 training dataset (Ardila et al., 2020) to prompt the CosyVoice2 synthesis process, thereby introducing speaker variability. Post-synthesis, we applied noise injection, random gain adjustments, and echo augmentation to simulate real-world acoustic conditions and enhance the model's adaptability to noisy environments.

Text-to-Text Retrieval

Speech-to-Speech and Text-to-Speech Retrieval In the speech-to-speech retrieval and Text-to-Speech scenario, our objective was to enable the model to directly learn matching relationships at the speech level. To this end, we adopted the SLUE-SQA-5 (Shon et al., 2023) and Spoken-SQuAD datasets (Li et al., 2018).

Audio+Text-to-Audio+Text Retrieval   For the more complex audio+text-to-audio+text retrieval task, which demands the integration of acoustic and semantic information, we curated and utilized six datasets: AudioSetSL, AudioCaps (Kim et al., 2019),MusicCaps (Agostinelli et al., 2023), Clotho (Drossos et al., 2019), VoxCeleb (Nagrani et al., 2017), and Xeno-canto (referred to as Xeno). In processing MusicCaps, AudioSetSL, and Clotho, we first computed audio similarity scores using the CLAP model. We then selected audio pairs which have the highest similarity to each other as positive matches. Subsequently,we feed these paired audio samples—along with their corresponding textual descriptions and predefined prompts—into the Gemini1.5 Pro model to generate extended knowledge, questions, and answers. Within each pair, one audio clip was designated as the "query audio," and the other as the "knowledge audio." The "knowledge audio" and its associated generated expanded knowledge served as contextual cues to guide the retrieval process for the "query audio" and the generated textual question. Consequently, the model is trained to retrieve the corresponding "knowledge audio" and expanded knowledge given a "query audio" and a textual question. For unpaired audio instances within the original datasets, we retained them for auxiliary tasks such as audio-to-text or text-to-audio retrieval, ensuring comprehensive data utilization. Furthermore, we implemented supplementary processing for the Xeno (bird species dataset) and VoxCeleb (celebrity voices dataset) datasets. In Xeno, we randomly selected samples per species and augmented them with species descriptions retrieved via the Google API. In VoxCeleb, we constructed paired samples for celebrity voices, supplementing them with biographical information obtained through Google API searches. We also proactively collected knowledge-intensive audio question-answering data, which often involves content with rich background knowledge, such as film production company intro music or diverse cover versions of famous songs. By associating these audio clips with relevant background information, we created audio-text pairs to further enhance the model's capacity to integrate complex acoustic scenes with external knowledge.

B.2 Prompt for Knowledge Extension

In this section,We show the entire prompt for our dataset's knowledge extension,question generation and answer in Figure 5.

B.3 Data Samples Display

This section presents several data samples from different datasets (see Tables 6, 7, 8, 9, and 10). Each sample is displayed in its own table formatted with two columns—one for the Field and one for the Content.

B.4 Representation Visualization

To visualize the embedding spaces learned by different models, we randomly selected 150 samples from our test dataset, representing the four retrieval scenarios. Each sample consisted of either an audio clip and its caption, or a group of paired audio-text data. We extracted embeddings for each sample using CLAP, the original Qwen2-Audio model, and our WavRAG retriever. These embeddings were then projected into a two-dimensional space using Principal Component Analysis (PCA). Figure 6 shows the resulting PCA visualizations. In contrast to CLAP and the original Qwen2-Audio, the WavRAG embeddings show no clear separation between modalities. Instead, the audio, text, and combined audio+text embeddings for a given piece of information are closely clustered, indicating that WavRAG consistently represents the same semantic content across modalities.

```
I need to build a scenario like this:
The user inputs an audio B and asks a question about its information
(such as paralinguistic information such as style, emotion or
characteristics, or the knowledge it represents). To assist in the
analysis, the system finds a similar audio A and its text description
in the knowledge base through audio similarity retrieval to assist
in answering.

I will give you two audio clips and their captions, You need to
complete the following tasks:

Knowledge Expansion
Based on the caption of Audio A, expand on professional and detailed
background information about the audio itself as well as its broader
context. The expansion should be factual and detailed — information
that a typical LLM may not readily have or answer with certainty —
and Knowledge must directly help answer the generated questions.
Write this expanded knowledge in a style similar to Wikipedia, ensuring
logical consistency. Integrate mentions of Audio B question's answer
naturally rather than inserting them forcefully. But it should not
explicitly mention the existence of audio B. Don't add too much
irrelevant information.

Question Generation
Generate a question about audio B, reflecting the need for audio
analysis. Cannot mention any information in Audio B in the question and
the questions should ensure that the LLM cannot accurately answer them
based on its own knowledge. The generated questions must be answerable
according to the knowledge of audio A above. Audio B and the question
should form a retrieval pair that uniquely retrieves audio A and its
knowledge. You need to ensure that the question reflects the advantages
of audio plus text dual-modal retrieval. Text alone cannot accurately retrieve.

Answer Generation
It must be an answer that can be answered based on the knowledge provided
by audio A and the knowledge that a typical large model may have. Do not
disclose the existence of Audio A to the user. When forming an answer,
use knowledge from Audio A in conjunction with the features of Audio B.
Ensure the final answer aligns with facts and audio B's caption but should
not explicitly mention the existence of audio B's caption.

You only need to output content in the following format:
Rewrite knowledge:

Question generation:

Answer generation:
```
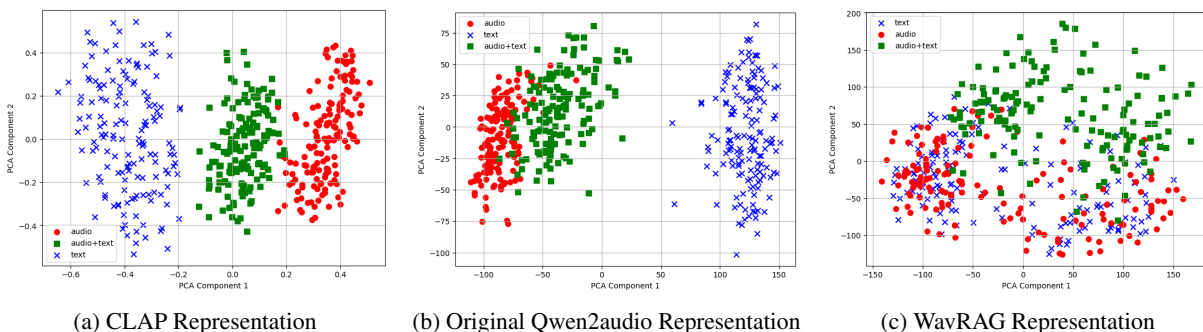
Figure 5: Prompt for extension

(a) CLAP Representation   (b) Original Qwen2audio Representation   (c) WavRAG Representation

Figure 6: Comparison of different model representations.

Task	Dataset		Instuction
Speech-to-Text	Quora		Given a question, retrieve questions that are semantically equivalent to the given question
	HotpotQA		Given a multi-hop question, retrieve documents that can help answer the question.
Text-to-Text	ELI5		Given a question, retrieve relevant documents that best answer the question.
	TrivialQA		Given a question, retrieve relevant documents that best answer the question.
	SQuAD		Given a question, retrieve relevant documents that best answer the question.
	MS MARCO		Given a web search query, retrieve relevant passages that answer the query.
Speech-to-Specch	SLUE-SQA-5		Please retrieve the most relevant speech in the document based on the following questions
Text-to-Speech	SpokenSQuAD		Based on the following text query, retrieve the most relevant speech.
Audio+Text-to-Audio+Text	Ours	AudioSetSL	Audio2Text:Based on the following audio, extract the most relevant text description
			Text2Audio:Based on the following text description, extract the most relevant audio
			AT2AT:Extract the most relevant knowledge based on the following questions
		AudioCaps	Audio2Text:Based on the following audio, extract the most relevant text description
			Text2Audio:Based on the following text description, extract the most relevant audio
			AT2AT:Extract the most relevant knowledge based on the following questions
		MusicCap	Audio2Text:Based on the following audio, extract the most relevant text description
			Text2Audio:Based on the following text description, extract the most relevant audio
			AT2AT:Extract the most relevant knowledge based on the following questions
		Clotho	Audio2Text:Based on the following audio, extract the most relevant text description
			Text2Audio:Based on the following text description, extract the most relevant audio
			AT2AT:Extract the most relevant knowledge based on the following questions
		VoxCeleb	Extract the most relevant knowledge based on the following questions
		Xeno-canto	Extract the most relevant knowledge based on the following questions
		Collected	Extract the most relevant knowledge based on the following questions

Table 5: Dataset-specific prompt

Field	Content
id_a	18
caption_a	Someone is playing a fast melody on a low bansuri flute along with someone playing tablas and a shrutibox in the background. This song may be playing at a live performance.
id_b	3031
caption_b	This is an Indian traditional music piece performed live by an orchestra. The orchestra is composed of the bansuri, the sitars and the tabla percussion. There is a vibrant atmosphere in the recording. This piece could be used in the soundtrack of a Bollywood movie.
rewrite_knowledge	The bansuri is a side-blown flute from the Indian subcontinent, made from a single hollow bamboo shaft with six or seven finger holes. An ancient instrument with deep cultural roots, it is associated with pastoral traditions, the love story of Krishna and Radha, and is revered in Hindu texts as Lord Krishna's divine instrument. The sound is generated by the resonance of the air column inside the flute, with pitch being controlled by covering the holes; half-holes are used to produce flat or minor notes. The bansuri's size determines its pitch: longer flutes produce a lower range, while shorter ones are higher. There are two main varieties: North Indian bansuri: A smaller, soprano instrument (approx. 14 inches) used for lighter compositions and film music. Bass bansuri: A larger version (approx. 30 inches) with a lower range, preferred for Hindustani classical music. In traditional performances, the bansuri is often accompanied by the tabla, a pair of twin hand drums that has been the main percussion instrument in Hindustani classical music since the 18th century. A continuous drone is provided by the shrutibox (or surpeti), a bellows-operated instrument. This combination of instruments is a hallmark of traditional Indian music, creating a rich, textured soundscape.
question_generation	What are the instruments used, and what is the style of the music being played?
answer_generation	The music features a traditional Indian ensemble, highlighted by the distinct sounds of the bansuri, sitar, and tabla. The bansuri, a type of bamboo flute, carries the melody with its characteristic reedy and airy tones, indicative of its use in both folk and classical Indian music. Accompanying the bansuri are the sitar, whose resonant strings add a layer of melodic complexity, and the tabla, providing a rhythmic foundation with its dynamic percussive beats. The interplay of these instruments suggests a live performance setting, capturing the vibrant essence of a traditional Indian music piece. This style is often associated with cultural and cinematic expressions, particularly in Bollywood.

Table 6: Musiccap Sample 1

Field	Content
id_a	34
caption_a	An acoustic drum is playing a simple rock groove with some drum fills on the toms and snare along with a distorted e-bass and e-guitar playing the same line. Panned to the right side of the speaker you can hear a piano sound playing a little melody. A female voice is singing a happy sounding melody while male backing voices are supporting her. This song may be playing sitting in your room enjoying being at home alone.
id_b	4585
caption_b	The low quality recording features a punk song that contains flat male vocal singing over punchy kick and snare hits, shimmering cymbals, wide electric guitars and groovy bass guitar. It sounds energetic, exciting and upbeat—like something you would jump to at concerts.
rewrite_knowledge	The audio features a typical rock ensemble, characteristic of the genre's instrumentation and arrangement during its peak popularity in the mid-20th century, particularly from the 1960s through the 1980s. The presence of an acoustic drum set laying down a "simple rock groove" indicates a 4/4 time signature with an emphasis on the backbeat (beats 2 and 4), a foundational element of rock music. The inclusion of drum fills on the toms and snare suggests dynamic variations within the song, a technique used to build tension and release or to transition between sections. The "distorted e-bass and e-guitar playing the same line" points to a common practice in rock music where the bass guitar doubles the guitar riff, creating a heavier, more unified sound. This technique is particularly prevalent in genres like hard rock and heavy metal, which emerged from the broader rock tradition. The distortion effect on the electric instruments is achieved through the use of overdrive or distortion pedals, or by cranking up the amplifier's gain, resulting in a "fuzzy" or "crunchy" tone that is synonymous with rock music. The "piano sound playing a little melody" panned to the right side adds a melodic counterpoint to the rhythm section. The use of panning creates a stereo image, giving the listener a sense of spatial depth. The female voice singing a "happy sounding melody" with male backing vocals suggests a lead and harmony vocal arrangement, a common feature in many rock subgenres.
question_generation	Considering the low quality and punk style of the recording, what is the likely era and subgenre of the rock music being played, and what musical characteristics might be expected in this context?
answer_generation	Based on the energetic, exciting, and upbeat nature of the low-quality recording, combined with the punchy drums, shimmering cymbals, wide electric guitars, and groovy bass, it sounds like a classic example of punk rock. This genre often features a raw, stripped-down sound, and the flat male vocals further reinforce this impression. The music's energy and the feeling it evokes, described as something you would "jump to at concerts," are hallmarks of punk's rebellious and lively spirit. Given the characteristics described, the song likely originates from the mid-1970s to early 1980s, a period when punk rock was flourishing and establishing its distinctive sound and ethos.

Table 7: Musiccap Sample 2

Field	Content
question_audio	blues/prideandjoy/Pride and Joy (Piano Cover).mp3
question	what song it is
answer	pride and joy
knowledge_audio	blues/prideandjoy/Stevie Ray Vaughan & Double Trouble - Pride and Joy (Official Audio).mp3
knowledge	"Pride and Joy" is a song by American singer, guitarist and songwriter Stevie Ray Vaughan and his backing band Double Trouble, released in late 1983 by Epic Records. It lists Vaughan as the writer, but actually it is rewritten from a 1962 record called "I Go Into Orbit" by Johnny Acey. The song was released on Stevie's debut studio album Texas Flood (1983). "Pride and Joy" was released as Vaughan's debut single and has become one of his most popular songs.

Table 8: Collected Sample

Field	Content
id	id10001
name	A.J._Buckley
intro	A.J. Buckley is an Irish-born Canadian actor best known for his role as Ed Zeddmore in the television series *Supernatural* and as Sonny Quinn in the military drama series *SEAL Team*. He has also appeared in numerous other television shows and films, often portraying tech-savvy characters or military personnel. Buckley is recognized for his intense on-screen presence and is also a co-founder of the Paperclip clothing line. He is married with three children.
audio_samples	J9lHsKG98U8/00016.wav; Y8hIVOBuels/00007.wav
question	Based on the voice, who do you think this is?
answer	A.J._Buckley

Table 9: Voxceleb Sample

Field	Content
id_a	7
filename_a	002_78_rpm_vinyl_noise_44_16_lossless.wav
captions_a	1. Static on a stereo or a similar device.
	2. White noise static is droning and crackling in the background.
	3. White noise static drones and crackles in the background.
id_b	2310
filename_b	Vinyl record noise.wav
captions_b	1. A record has reached the end of a song and is waiting to start over.
	2. A turntable turns with no needle touching record.
	3. An electronic hiss and pop is all the speaker emits.
	4. The cracking of a small fire very close to the receiver.
	5. The crackling of a small fire very close to the receiver.
rewrite_knowledge	Static in audio refers to a random hissing or crackling sound, often perceived as white noise. In audio equipment like stereos, it can originate from atmospheric interference affecting radio reception, or from internal electronic issues like faulty components or poor shielding. When heard from a stereo's speakers, this often indicates a problem within the amplifier or circuitry and is distinct from the intentional use of static for artistic effect
question_generation	What sound is it?
answer_generation	The sound you're hearing likely originates from a mechanical audio playback device, possibly one utilizing a physical medium for sound storage. The pause and the nature of the sound suggest the equipment has reached the end of a discrete segment of audio, such as a track on a record. Older audio technologies often employed physical mechanisms to play and transition between recordings. The sound is characteristic of a system waiting for a manual or automatic command to initiate the next segment of audio playback, a process that is distinct from the seamless transitions found in modern digital audio systems. This pause, accompanied by mechanical sounds, is a hallmark of certain analog audio formats where playback is not continuous but segmented by the physical structure of the recording medium.

Table 10: Clotho Sample