FEEDBACK-DRIVEN RETRIEVAL-AUGMENTED AUDIO GENERATION WITH LARGE AUDIO LANGUAGE MODELS

Junqi Zhao[1,2], Chenxing Li[2], Jinzheng Zhao[1,2], Rilin Chen[2],
Dong Yu[3], Mark D. Plumbley[1], Wenwu Wang[1,*]

[1] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
[2] Tencent AI Lab, Beijing, China
[3] Tencent AI Lab, Seattle, USA

ABSTRACT

We propose a general feedback-driven retrieval-augmented generation (RAG) approach that leverages Large Audio Language Models (LALMs) to address the missing or imperfect synthesis of specific sound events in text-to-audio (TTA) generation. Unlike previous RAG-based TTA methods that typically train specialized models from scratch, we utilize LALMs to analyze audio generation outputs, retrieve concepts that pre-trained models struggle to generate from an external database, and incorporate the retrieved information into the generation process. Experimental results show that our method not only enhances the ability of LALMs to identify missing sound events but also delivers improvements across different models, outperforming existing RAG-specialized approaches.

Index Terms— Text-to-audio generation, retrieval augmented generation, large audio-language models

1. INTRODUCTION

Text-to-audio (TTA) generation refers to the task of generating audio content conditioned on natural language descriptions. This field has been transformed by recent advances in generative models, especially diffusion-based [1, 2, 3] and language model-based approaches [4, 5, 6].

These generative models require extensive training data and computational resources. Due to their reliance on the training data, they struggle to generate audio categories that are rare or unseen in the dataset [7], along with user-specified audio content. This limitation substantially hinders their generative performance.

To address the class-imbalance problem in the training data of TTA models, Re-AudioLDM [7] builds upon the AudioLDM [1] model and retrieves relevant audio-text pairs from the database as supplementary information, thereby improving the modeling of certain rare sound events during training. Audiobox TTA-RAG [8], a recent retrieval-

augmented TTA method based on Audiobox [9], extends the flow-matching model by augmenting the conditioning input with both text prompts and retrieved audio samples, and eliminates the need for labels in the external audio data source. However, these approaches necessitate training a dedicated RAG-assisted Latent Diffusion Model (LDM) or Flow Matching (FM) model from scratch. This significantly increases the training cost, and whenever a new concept must be generated, the model's architecture upgraded, or the database updated, the model needs to be retrained.

In contrast to prior RAG-specialized TTA models, we propose a novel feedback-guided retrieval augmentation approach built upon a Large Audio Language Model (LALM) to address the limitations of current TTA models, particularly their suboptimal generation performance and the omission of certain sound events. Specifically, we first use the fine-tuned LALM [10] to evaluate the audio generated by the pre-trained TTA model, assessing whether it aligns with the text prompt and identifying any missing or suboptimal sound events. Next, for audio concepts that the model struggles to generate effectively, we retrieve relevant samples from an external database. Finally, we employ a decoupled cross-attention mechanism to integrate the retrieved audio as references into the pre-trained TTA model during inference, steering the model toward producing the desired audio outputs and thereby enhancing its overall generation efficacy.

We conducted experiments on different pre-trained TTA models, including both diffusion-based [11] and flow-matching-based [3] models. We will show that our method can be applied across all pre-trained TTA models, without the need to train a dedicated retrieval-augmented TTA model from scratch as required by approaches such as Re-AudioLDM or Audiobox TTA-RAG, thereby ensuring broad applicability. Our proposed method achieves substantial performance improvements across different TTA models, outperforming the state-of-the-art (SOTA) RAG-based TTA approaches by a significant margin.

Our main contributions can be summarized as follows: 1) We introduce a feedback-driven RAG method based on an
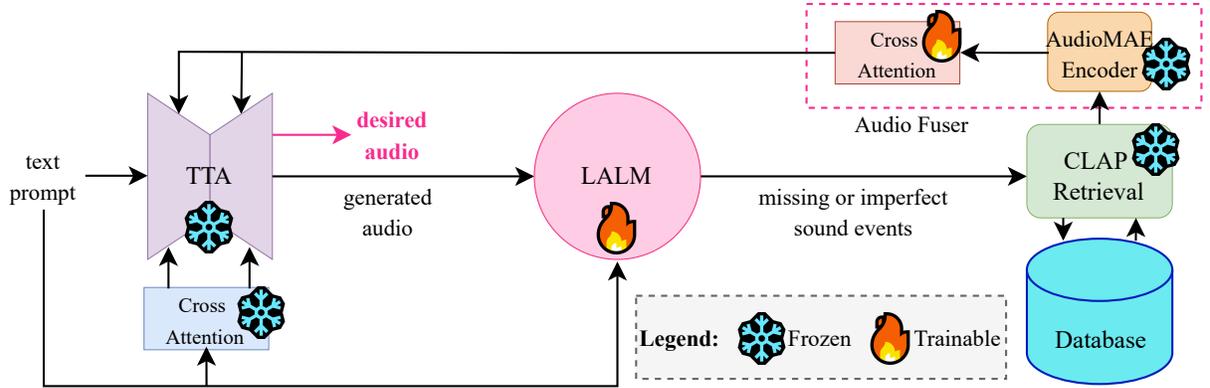
---

*Corresponding author.

Fig. 1. The overview structure of our proposed method.

LALM to enhance the performance of pre-trained TTA models. 2) To improve the accuracy of the LALM in identifying missing sound events, we further fine-tune it using Low-Rank Adaptation (LoRA). 3) To tackle the challenges faced by current TTA models in complex multi-event audio scenes, we develop a lightweight audio fuser with decoupled cross-attention, which enhances audio events that are missing or poorly generated during inference.

2. RELEATED WORK

Retrieval-augmented generation, first introduced in natural language processing (NLP) [12, 13], has emerged as a promising approach to mitigate hallucination and outdated knowledge in large language models (LLMs). Subsequently, RAG has attracted increasing research attention in the field of image generation [14, 15, 16, 17]. For speech generation tasks, Xue et al. [18] introduced a Context-Aware Contrastive Language-Audio Pre-training model that captures style-related contextual features and leverages contextual information in the retrieval process to improve the selection of specialized audio prompts. AutoStyle-TTS [19] presents a text-to-speech framework leveraging RAG, which dynamically adapts speech styles to textual context, thereby enabling more natural and expressive speech synthesis.

For the TTA generation task examined in this study, Re-AudioLDM [7] and Audiobox TTA-RAG [8] attribute the performance limitations to class imbalance in the training datasets and therefore train a dedicated RAG-based model from scratch on top of a base TTA model.

3. PROPOSED METHOD

As shown in Figure 1, our proposed method consists of three main components: a finetuned LALM module that evaluates the outputs of pre-trained TTA models; a retrieval module that searches the database for relevant audio using missing or im-

perfect sound event captions generated by the LALM; and a TTA model enhanced with an audio fuser.

The workflow of our system proceeds as follows: we begin by augmenting the pretrained TTA model with a decoupled cross-attention module for audio, yielding an audio-enhanced TTA model. During the retrieval stage, the parameters of the audio-enhanced TTA model remain fixed. For the currently popular pre-trained TTA models, the input is the textual description of the target audio and the output is the corresponding audio clip. However, when faced with complex audio scenes or rare and even unseen audio types, the model may only generate a subset of the intended sound events, or the quality of certain generated events may be unsatisfactory. A fine-tuned LALM can interpret the audio outputs of the pre-trained TTA model and identify sound events that are either missing or poorly generated, thereby enabling targeted enhancement. This enhancement is achieved by retrieving the corresponding audio from an external database and feeding it back to the pre-trained TTA model through our designed audio fuser.

3.1. Large Audio Language Model

Qwen2.5-Omni [10] represents a SOTA LALM capable of simultaneously processing auditory and textual information and producing real-time outputs in both text and natural speech. It adopts a Thinker-Talker architecture.

Although Qwen2.5-Omni demonstrates strong performance across various audio understanding benchmarks, its effectiveness is limited in the missing or imperfect sound event recognition task required in this study. To address this issue, we fine-tune the Qwen2.5-Omni-7B model on our task-specific dataset using the LoRA approach. More concretely, we build a dataset for the task of missing sound event identification. Each instance consists of an audio sample for the LALM to interpret and a text query: "What sound events are missing from the audio compared to `<target prompt>`?" The LALM is required to output the sound events that are not

present in the audio but appear in the target prompt.

3.2. Retrieval Module

Similar to Re-AudioLDM and Audiobox TTA-RAG, we employ the Contrastive Language–Audio Pretraining (CLAP) model as the retrieval module to query the database for audio samples based on the text descriptions of missing or imperfect sound events. Since our approach adopts text-to-audio retrieval, it does not require labeled external data sources, thereby supporting large-scale retrieval from in-the-wild and unlabeled audio datasets.

3.3. Audio-Enhanced Text-to-Audio Generation

The audio-enhanced TTA model consists of two components: an original pre-trained TTA model and an audio fuser. The audio fuser is designed to enable the pre-trained TTA model to generate audio conditioned on retrieved audio prompts. To allow the TTA model to capture fine-grained features from the audio prompt, we first employ an Audio Masked Autoencoder (MAE) [20] to extract representations from the input audio. We then introduce an adapted module based on a decoupled cross-attention mechanism [21], which separates the cross-attention layers for audio and text, in order to inject these features into the internal layers of the pre-trained TTA model.

In the original pre-trained TTA, the text encoder features c_t are injected into a diffusion model through cross-attention layers:

$$z_t = \text{Attention}\big(zW_q,\ c_t W_k,\ c_t W_v\big). \quad (1)$$

Here, z denotes the internal feature of the diffusion model; $Q = zW_q$, $K = c_t W_k$, and $V = c_t W_v$ correspond to the query, key, and value in the attention operation, respectively, where W_q, W_k, and W_v are the weight matrices of the linear projection layers.

Instead of inserting audio features by concatenating them with text features and feeding the result into the cross-attention layers, we add an additional cross-attention layer to each cross-attention layer in the original TTA model to integrate audio features. Given the AudioMAE features c_a, the output of the new cross-attention is computed as follows:

$$z_a = \text{Attention}\big(zW_q,\ c_a W_k',\ c_a W_v'\big), \quad (2)$$

where W_k' and W_v' are the newly added weight matrices. Then, we add the output of the audio cross-attention to that of the text cross-attention via:

$$z^{\text{new}} = z_t + \lambda z_a \quad (3)$$

where λ is a weight factor that controls the strength of the retrieved audio.

In the training phase of the audio-enhanced TTA model, the parameters of the original TTA model are frozen, while only W_k' and W_v' are updated, enabling parameter-efficient fine-tuning and reducing training overhead.

Model	BERTScore (%)	SimCSE (%)
Ground Truth	100.0	100.0
Gemini 2.5 Pro	80.2	89.1
Qwen2.5-Omni-7B	53.8	73.8
Qwen2.5-Omni-7B (SFT)	**93.3**	**92.6**

Table 1. Comparison of different models for missing sound event identification.

4. EXPERIMENTS

4.1. Datasets

We constructed a training dataset for the missing sound event identification task using the AudioCaps (AC) training dataset and the AudioSet (AS) balanced training dataset, comprising a total of 45,222 samples. Similarly, we construct the test set using the AudioCaps test set and the AudioSet evaluation set.

For training the audio-enhanced TTA model, we only train the cross-attention layers for audio using the AC training set. The external database consists of the AS balanced subset and the Freesound dataset. For RAG evaluation, we employ the AC test set along with the RiTTA [22] Count test set.

4.2. Experimental Setup

Implementation Details. For Qwen2.5-Omni, we fine-tune the 7B version of the model with a LoRA rank of 8. The model is trained for 5 epochs with a batch size of 4. As for the audio-enhanced TTA model, we employ two base models, AudioLDM2-Large [11] and TangoFlux [3]. All parameters in the pre-trained TTA model and the AudioMAE are frozen, except for the decoupled audio cross-attention layers. Training is conducted for 20,000 steps with an effective batch size of 28. We adopt the AdamW optimizer with a fixed learning rate of 1×10^{-4} and a weight decay of 1×10^{-2}. To enable classifier-free guidance at inference, we randomly drop audio and text conditions with a probability of 5% during training.

Evaluation Metrics. Following Re-AudioLDM and Audiobox TTA-RAG, we evaluate the effectiveness of our proposed method using Fréchet Distance (FD), Kullback–Leibler (KL) divergence, Inception Score (IS), and CLAP score. Lower FD and KL scores reflect higher audio quality, while a higher IS score indicates better quality and greater diversity in the generated audio. The CLAP score measures the degree of semantic alignment between the generated audio and the given text.

In addition, to evaluate the accuracy of the LALM in the missing sound event identification task, we adopt BERTScore [23] and SimCSE [24] as evaluation metrics, as they are important measures of textual similarity in NLP.

Model	Dataset	Retrieval Info.	Database & Retrieval No.	KL ↓	FD ↓	IS ↑	CLAP (%) ↑
Re-AudioLDM-L	AudioCaps	Audio & Text	AC →10	**1.20**	–	7.39	37.12
Audiobox TTA-RAG	AudioCaps	Audio	AC →3	1.44	–	8.40	37.37
AudioLDM2	AC+AS+6 others	✗	✗	1.59	33.2	7.40	45.20
AudioLDM2-RAG (ours)	AC+AS+6 others	Audio	AS →1	1.55	30.6	8.49	46.22
TangoFlux	AC+1 other	✗	✗	1.21	19.23	12.60	**58.60**
TangoFlux-RAG (ours)	AC+1 other	Audio	AS →1	**1.20**	18.98	12.81	**58.60**

Table 2. Performance comparison of different models on the AudioCaps test set (ID).

5. RESULTS

5.1. Missing Sound Events Identification

In our experiments, we observed that Qwen2.5-Omni tends to generate hallucinations in the task of identifying missing sound events. To address this issue, we performed supervised fine-tuning (SFT) on our custom dataset. The experimental results of different LALMs on the test set are presented in Table 1. The results indicate that although Qwen2.5-Omni-7B achieves state-of-the-art performance on many common audio understanding tasks, its zero-shot performance on the missing sound event identification task remains unsatisfactory. With supervised fine-tuning, however, the performance of Qwen2.5-Omni-7B (SFT) improves substantially and even surpasses that of Gemini 2.5 Pro. This fine-tuning enables us to leverage LALMs to more effectively evaluate the outputs of pre-trained TTA models in identifying missing or low-quality sound events, thereby facilitating the retrieval of relevant audio from external databases to enhance generation performance.

5.2. Main Results

The experiments are conducted on two test sets: the Audio-Caps test set and the RiTTA Count test set, which represent in-distribution (ID) results and out-of-distribution (OOD) generalization results, respectively. We compare our performance against several state-of-the-art models, including Re-AudioLDM [7], Audiobox TTA-RAG [8], AudioLDM2 [11], and TangoFlux [3].

5.2.1. RAG Results on AudioCaps

The generation performance of different models on the AudioCaps dataset is presented in Table 2. We use the AudioSet balanced subset as our database. When applying our proposed RAG method, both AudioLDM2-RAG and TangFlux-RAG exhibit notable performance improvements over their original TTA models. This demonstrates that our approach is highly effective across different base models. Moreover, it eliminates the need to train a dedicated RAG model from

Model	KL ↓	FD ↓	FAD ↓	IS ↑	CLAP (%) ↑
AudioLDM2	2.81	38.5	7.7	7.4	29.0
AudioLDM2-RAG (ours)	**2.71**	**35.2**	**4.4**	**8.5**	**34.2**
TangoFlux	2.22	46.8	7.3	7.0	43.3
TangoFlux-RAG (ours)	**2.18**	**37.7**	**5.1**	**7.3**	**43.7**

Table 3. Performance comparison of different models on the RiTTA Count test set (OOD).

scratch, significantly reducing the training costs. In addition, our TangoFlux-RAG achieves the best performance compared with the baseline models across all evaluation metrics. The improvement over TangoFlux is not very large, which may be due to the fact that TangoFlux is currently the strongest TTA model, and that the training and test distributions are highly similar, which may introduce evaluation bias.

5.2.2. RAG Results on RiTTA

We selected the RiTTA Count test set for out-of-distribution evaluation, with the AudioSet balanced subset and the Freesound dataset as the database. Table 3 reports the evaluation results of the base model and our RAG method. This experiment demonstrates that when the distributions of the training and test data differ, our method achieves better performance on the RiTTA test set than its base model. Overall, the poorer the model's performance on a given metric, the greater the improvement obtained with our method.

6. CONCLUSION

In this work, we present a general RAG method that can be applied to various pretrained TTA models and databases, eliminating the need to train a specialized RAG-based TTA model from scratch. To improve performance, we utilize an LALM to evaluate the outputs of the base model, and then introduce a lightweight audio fuser to incorporate the retrieved audio. Experiments show that our method delivers consistent performance improvements in both in-domain and out-of-domain settings, outperforming dedicated RAG models.

7. REFERENCES

[1] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 21450–21474.

[2] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction guided latent diffusion model," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3590–3598.

[3] C.-Y. Hung, N. Majumder, Z. Kong, A. Mehrish, A. A. Bagherzadeh, C. Li, R. Valle, B. Catanzaro, and S. Poria, "TangoFlux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization," *arXiv preprint arXiv:2412.21037*, 2024.

[4] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "AudioGen: Textually guided audio generation," in *The Eleventh International Conference on Learning Representations*.

[5] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, X. Chang, J. Shi, S. Zhao, J. Bian, X. Wu *et al.*, "UniAudio: An audio foundation model toward universal audio generation," *arXiv preprint arXiv:2310.00704*, 2023.

[6] S.-w. Yang, B. Kim, K.-P. Huang, Q. Tang, H. PHAN, B.-R. Lu, H. Sundar, S. Ghosh, H.-y. Lee, C.-C. Kao *et al.*, "Generative audio language modeling with continuous-valued tokens and masked next-token prediction," in *Forty-second International Conference on Machine Learning*.

[7] Y. Yuan, H. Liu, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Retrieval-augmented text-to-audio generation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2024, pp. 581–585.

[8] M. Yang, B. Shi, M. Le, W.-N. Hsu, and A. Tjandra, "Audiobox TTA-RAG: Improving zero-shot and few-shot text-to-audio with retrieval-augmented generation," *arXiv preprint arXiv:2411.05141*, 2024.

[9] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan *et al.*, "Audiobox: Unified audio generation with natural language prompts," *arXiv preprint arXiv:2312.15821*, 2023.

[10] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang *et al.*, "Qwen2.5-Omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[11] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[12] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[13] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, vol. 2, no. 1, 2023.

[14] A. Blattmann, R. Rombach, K. Oktay, J. Müller, and B. Ommer, "Retrieval-augmented diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 15309–15324, 2022.

[15] S. Sheynin, O. Ashual, A. Polyak, U. Singer, O. Gafni, E. Nachmani, and Y. Taigman, "KNN-Diffusion: Image generation via large-scale retrieval," in *The Eleventh International Conference on Learning Representations*.

[16] W. Chen, H. Hu, C. Saharia, and W. W. Cohen, "Re-Imagen: Retrieval-augmented text-to-image generator," in *The Eleventh International Conference on Learning Representations*.

[17] R. Shalev-Arkushin, R. Gal, A. H. Bermano, and O. Fried, "ImageRAG: Dynamic image retrieval for reference-guided image generation," *arXiv preprint arXiv:2502.09411*, 2025.

[18] J. Xue, Y. Deng, Y. Gao, and Y. Li, "Retrieval augmented generation in prompt-based text-to-speech synthesis with context-aware contrastive language-audio pretraining," in *Proc. Interspeech 2024*, 2024, pp. 1800–1804.

[19] D. Luo, C. Ma, W. Li, J. Wang, W. Chen, and Z. Wu, "AutoStyle-TTS: Retrieval-augmented generation based automatic style matching text-to-speech synthesis," *arXiv preprint arXiv:2504.10309*, 2025.

[20] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," *Advances in Neural Information Processing Systems*, vol. 35, pp. 28708–28720, 2022.

[21] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[22] Y. He, Y. Jain, X. Liu, A. Markham, and V. Vineet, "RiTTA: Modeling event relations in text-to-audio generation," *arXiv preprint arXiv:2412.15922*, 2024.

[23] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with bert," in *International Conference on Learning Representations*.

[24] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 6894–6910.