Text Prompt is Not Enough: Sound Event Enhanced Prompt Adapter for Target Style Audio Generation

Chenxu Xiong[1], **Ruibo Fu**[2,*], **Shuchen Shi**[2], **Zhengqi Wen**[3], **Tao Wang**[2], **Chenxing Li**[4], **Chunyu Qiang**[2], **Yuankun Xie**[2], **Xin Qi**[2], **Guanjun Li**[2], **Zizheng Yang**[1]
[1] SDU-ANU Joint Science College, Shandong University, Weihai, China
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3] Department of Automation, Tsinghua University, Beijing, China [4]
AI Lab, Tencent, Beijing, China
`202100700062@mail.sdu.edu.cn, ruibo.fu@nlpr.ia.ac.cn`

Abstract

Current mainstream audio generation methods primarily rely on simple text prompts, often failing to capture the nuanced details necessary for multi-style audio generation. To address this limitation, the Sound Event Enhanced Prompt Adapter is proposed. Unlike traditional static global style transfer, this method extracts style embedding through cross-attention between text and reference audio for adaptive style control. Adaptive layer normalization is then utilized to enhance the model's capacity to express multiple styles. Additionally, the Sound Event Reference Style Transfer Dataset (SERST) is introduced for the proposed target style audio generation task, enabling dual-prompt audio generation using both text and audio references. Experimental results demonstrate the robustness of the model, achieving state-of-the-art Fréchet Distance of 26.94 and KL Divergence of 1.82, surpassing Tango, AudioLDM, and AudioGen. Furthermore, the generated audio shows high similarity to its corresponding audio reference. The demo, code, and dataset are publicly available.[2]

1 Introduction

Target Style Audio Generation generates audio with specific styles or features, allowing for more natural and fine-grained audio production. This approach has numerous applications, particularly in the media industries, where it can generate background sound effects that match specific scenes. The current mainstream method for audio generation is Text-to-Audio (TTA) [Yang et al., 2023][Kreuk et al., 2023][Huang et al., 2023][Liu et al., 2023][Ghosal et al., 2023]. These TTA models, often encoded by CLAP [Elizalde et al., 2022] or T5 [Raffel et al., 2019], utilize rich semantic information in textual descriptions to produce high-quality audio outputs.

Although mainstream methods using single-text prompts have achieved promising results, several limitations remain. Text input and audio output belong to different modalities, making alignment between the two challenging. For instance, generating the sound of a dog barking from a single text prompt fails to capture specific characteristics such as timbre or how the environment interacts with the barking. This limitation restricts the ability to model audio in finer detail. To address this issue, incorporating additional prior knowledge is essential for providing richer contextual information and enhancing the precision of the generated output.

* *Corresponding author
[2] https://michael1223132.github.io/PromptAdapter/

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Two primary approaches exist for introducing prior knowledge into audio generation. The first involves control conditions manipulating the generated audio's pitch, energy, and temporal relationships [Guo et al., 2023][Xie et al., 2024][Liao et al., 2024]. However, no current methods specifically address style control in audio generation. The second approach utilizes multi-modal prompts that incorporate semantic and temporal information from other modalities, such as images [Sheffer and Adi, 2022] and videos [Iashin and Rahtu, 2021][Luo et al., 2023][Xu et al., 2024]. Despite their potential, cross-modal prompts often suffer from interference caused by redundant and unrelated information, as they do not provide intuitive acoustic references for the model.

In this paper, we first propose the Sound Event Enhanced Prompt Adapter. Traditional style transfer approaches typically extract a global style directly from the reference. However, text offers valuable semantic information that can guide and refine the application of this global style. To leverage this, cross-attention [Vaswani et al., 2017] is employed between sound events and text to identify which text events are most closely correlated with the corresponding audio reference. Additionally, the style embedding generated by the adapter is passed into the U-Net [Ronneberger et al., 2015] via adaptive layer normalization [Peebles and Xie, 2022], which allows the normalization layer to adapt to the data distribution from style embedding. We then construct a Sound Event Reference Style Transfer Dataset (SERST) that integrates dual-modality prompts from event-level audio reference and text, derived from Audioset-Strong [Hershey et al., 2021]. Experimental results demonstrate the robustness of the proposed method across various sound event references, and significant improvements in the accuracy of acoustic modeling. Specifically, the method achieves gains of 2.3% in Fréchet Distance and 7.6% in KL divergence. Additionally, the generated audio exhibits a strong alignment with its audio reference, as indicated by a score of 0.4 in CLAP-audio similarity. The key contributions of this paper are summarized as follows:

- A new audio generation task is introduced, guided by both text and sound event references, enabling the transfer of style from the reference and improving the accuracy and naturalness of audio generation.
- A new dataset, SERST, is created by integrating existing datasets, consisting of audio and sound event segments. Evaluation metrics were applied to assess performance, providing a benchmark for future research.
- A Sound Event Enhanced Prompt Adapter is proposed that adaptively transfers the style from reference audio through cross-attention between the text and reference audio segments, integrated with an adaptive layer normalization within the U-Net. This approach enables finer-grained control over the audio generation process that enhances the accuracy of acoustic modeling and achieves target style transfer.

2 Method

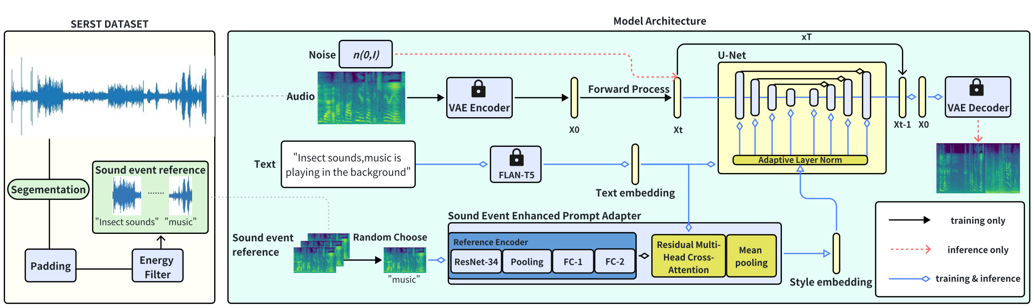

Figure 1: SEARST Dataset and Model Architecture: In the training stage, the latent diffusion model (LDM) is conditioned on an audio embedding learned in a continuous space through a variational auto-encoder (VAE). Text is fused with a randomly selected sound event reference through the Sound Event Enhanced Prompt Adapter to generate a style embedding. This style embedding is then utilized for adaptive layer normalization in the U-Net. In the inference stage, the LDM is conditioned on random noise instead of the audio embedding derived from the VAE.

2.1 Sound Event Reference Style Transfer Dataset (SERST)

Effective style transfer requires high-quality reference audio. To address this need, the Sound Event Reference Style Transfer Dataset (SERST) is constructed, providing event-level granularity audio to capture the full distribution of acoustic events and enabling the accurate reconstruction of their characteristics. This dataset is created by segmenting the original audio from the Audioset-Strong dataset [Hershey et al., 2021] based on annotated acoustic event timestamps. Statistical analysis revealed that a 2-second audio length offers an optimal balance between segment quantity and accuracy. Audio is segmented by event, and in cases where the segments are shorter than 2 seconds, they are concatenated from other clips with the same sound event tag, facilitating both padding and data augmentation. Then a Short-Time Energy detection is used to filter out Poor quality references. As a single original audio could yield multiple trimmed segments: during training, one of these segments is randomly selected, while during inference, all trimmed segments are utilized to examine the variability in the generated audio. The dataset consists of 88,464 training samples, 1,384 validation samples, and 1,180 test samples.

2.2 Sound Event Enhanced Prompt Adapter

To fully utilize the acoustic information, the global sound event style feature is extracted from a reference encoder. A style embedding is then generated through cross-attention between the text and reference audio, enabling adaptive style transfer and allowing the model to focus on the relevant aspects of the reference audio's style.

The sound event reference is first compressed into a reference embedding e_{r}, representing the global style of the audio. Given the lack of suitable pre-trained encoders for this task, a custom reference audio encoder was developed based on the H/ASP model [Heo et al., 2020], originally designed for Text-to-Speech (TTS). The global style is then integrated with local information from the text condition e_{t}. Residual cross-attention between the text embedding and the audio embedding is applied to generate the style embedding e_s:

$$Q = e_t W_q, \quad K = e_r W_k, \quad V = e_r W_v, \tag{1}$$

$$e_s = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d/h}}\right) \cdot V + e_t \tag{2}$$

d represents the embedding dimension of e_{t}, and h refers to the number of multi-heads. We then perform mean pooling along the sequence length dimension to align dimensions and feed them into U-Net.

2.3 Conditional Latent Diffusion Audio Generation Model

The LDM model aims to conduct the denoising process on mel-embedding (training) or standard Gaussian noise ϵ (inference) and predict the mel-embedding $\hat{x_0}$. For every step t, the training objective is to minimize the following:

$$\mathcal{L}_{\mathrm{LDM}} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,I)} \left\|\epsilon_\theta(x_t, t, e_{\mathrm{t}}, e_{\mathrm{r}}) - \epsilon\right\|_2^2 . \tag{3}$$

In this context, ϵ_θ represents the noise estimation conditioned on t, e_t and e_r. The architecture of the LDM primarily utilizes a U-Net structure [Ronneberger et al., 2015], which consists of a series of ResNet[He et al., 2015] and transformer blocks.

The shift parameters γ and β, derived from the concat of style embedding and time step embedding, are applied as adaptive layer normalization-zero parameters [Peebles and Xie, 2022] throughout the Resnet blocks in U-Net. This is because the adaptive layer norm allows the normalization layer to adapt to data distributions in different modalities or domains, thus performing well in multimodal learning or domain adaptation tasks.

3 Expiriments

3.1 Training Setting

All data are resampled to a 16kHz sampling rate, with each sample padded to a duration of 10.24 seconds. The VAE and text condition encoder are kept frozen and accept audio at 16kHz while we

Table 1: Sensitivity Analysis. The results show the Clap similarity of our generated audios under identical sound event reference (ID ref) or different sound (Diff ref) event reference.

Ours	***ID ref***	***Diff ref***	***Diff***
CLAP-Audio	0.72	0.54	**0.18**

Table 2: Model effectiveness. The results of the model effectiveness show the accuracy of generated audio from Our model compared to different baseline models.

Models	**Objective Metrics**			**Subjective Metrics**	
	FD ↓	**FAD** ↓	**KL** ↓	**OVL** ↑	**REL** ↑
Ground truth	–	–	–	87.50	83.65
AudioGen [Kreuk et al., 2023]	28.52	2.47	2.12	73.25	71.90
AudioLDM [Liu et al., 2023]	28.07	2.44	2.01	72.60	69.85
Tango [Ghosal et al., 2023]	27.60	**2.21**	1.97	74.40	75.40
Ours	**26.94**	2.38	**1.82**	**79.10**	**77.65**

fine-tuned the latent diffusion model using pre-trained weights from Tango [Ghosal et al., 2023]. The reference audio encoder is trained from scratch. The text encoder is based on FLAN-T5-LARGE [Chung et al., 2022], which contains a total of 780 million parameters. HiFi-GAN [Kong et al., 2020] is used as the vocoder. The trainable components include the U-Net, which loaded the pre-trained weights from Tango, and the reference audio encoder, collectively comprising 1.097 billion trainable parameters. Our model was trained for 20 epochs on four RTX 3090 GPUs with a batch size of two.

3.2 Evaluation Metrics

We compared our model to Tango[Ghosal et al., 2023], AudioGen [Kreuk et al., 2023] and AudioLDM[Liu et al., 2023] and used four objective metrics: Fréchet Distance (**FD**), Fréchet Audio Distance (**FAD**), KL divergence (**KL**), Mel-Spectrogram cosine Similarity (**Mel-Sim**) and CLAP-Audio[Elizalde et al., 2022] cosine similarity (**CLAP-Audio**).

As for subjective evaluation, we paid twenty experienced human evaluators to assess fifty randomly selected audio samples on a scale from 1 to 100 in the following aspects: overall audio quality (**OVL**) and relevance to the input text (**REL**) that reflects the quality of generated audio and its relevance to the input sound event prompt (**REA**) that demonstrates the ability in target style transfer.

4 Results and Analysis

4.1 Sensitivity Analysis for Sound Enhanced Prompt Adapter

Table 1 presents the CLAP-Audio similarity results of the generated audio provided with various sound event references, while keeping the text input constant. When the same sound event reference is provided to the model multiple times, the generated audio exhibits a CLAP similarity score of 0.72. In contrast, when different sound event references are used, the generated outputs yield a CLAP similarity score of 0.54. This difference of 0.18 demonstrates the effectiveness of the Sound Enhanced Prompt Adapter in utilizing prior acoustic information.

4.2 Comparison of Generated Audio Accuracy with Baseline Models

Table 2 presents the evaluation results of our model compared to TTA models using both objective and subjective metrics. In terms of objective metrics, our model achieves an FD score of 26.94, and a KL divergence of 1.88, which are all the lowest in all models. The FAD score of 2.38, although not the best, is still very competitive. For subjective metrics, our model achieves an OVL score of 79.10 and a REL score of 77.65, which are both the best in these models, showing that the audio generated by our model is very well aligned with the provided textual descriptions.

Table 3: Ablation study. The input channel implies where the style embedding will be sent into U-net after fusion. Fusion type means how we fuse text with reference.

Model	**Fusion Method**		**Objective Metrics**		
	Input Channel	**Fusion Type**	**FD** ↓	**FAD** ↓	**KL** ↓
Ours	**Timestep**	**Cross Attention**	**26.94**	**2.38**	**1.88**
Variant1	Timestep	Concat	28.54	3.14	1.93
Variant2	Text	Cross Attention	39.15	6.09	2.27
Variant3	Text	Concat	38.50	4.35	2.30

Table 4: Audio Relevance Evaluation. The results emphasize the alignment between the generated audio and its reference.

Models	**Mel-Sim** ↑	**CLAP-Audio**↑	**REA** ↑
AudioGen [Kreuk et al., 2023]	0.71	0.33	63.35
AudioLDM [Liu et al., 2023]	0.70	0.32	64.00
Tango [Ghosal et al., 2023]	0.71	0.34	64.25
Variant1	0.73	0.36	69.00
Ours	**0.76**	**0.40**	**76.00**

4.3 Ablation Study of Text and Sound Event Prompt Fusion Methods

Table 3 presents the results of our ablation study. We experimented with four different approaches: concatenating the reference embedding with the text embedding or applying cross-attention to obtain the style embedding, then sending the merged style embedding into U-net either with the text input or integrating it into the layer normalization of ResNet blocks within the U-Net, alongside the timestep embedding. The results indicate that using cross-attention to generate the style embedding, followed by its incorporation into the layer normalization, yields the best performance.

4.4 Style transfer performance evaluation by measuring audio similarity

Tabel 4 presents the evaluation results for the similarity of generated audio and sound event reference. Our model achieves the highest scores in all metrics, with a Mel-Sim of 0.76, CLAP-Audio similarity of 0.40, and an REA of 76.00, demonstrating strong relevance with the reference compared to the other models. These results underscore the effectiveness of our approach in leveraging sound event reference to transfer the generated audio.

5 Conclusion

This work first introduces the SERST dataset, which integrates dual-modality prompts from event-level audio reference and text, providing a valuable resource for target audio generation. Then a Sound Event Enhanced Prompt Adapter is proposed to achieve fine-grained style control in audio generation. The method leverages cross-attention and adaptive layer normalization, significantly improving the quality and controllability of generated audio, particularly in style. Compared to Tango, the proposed approach improves FD and KL Divergence scores by 2.3% and 7.6%. The generated audio strongly aligns with the reference audio, highlighting effective style control.

6 Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) (No.62101553).

References

H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, D. Valter, S. Narang, G. Mishra, A. W. Yu, V. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. hsin Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416, 2022.

B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang. Clap: Learning audio concepts from natural language supervision, 2022.

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*, pages 3590–3598, 2023.

Z. Guo, J. Mao, R. Tao, L. Yan, K. Ouchi, H. Liu, and X. Wang. Audio generation with multiple conditional diffusion model. In *AAAI Conference on Artificial Intelligence*, 2023.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

H.-S. Heo, B.-J. Lee, J. Huh, and J. S. Chung. Clova baseline system for the voxceleb speaker recognition challenge 2020. *ArXiv*, abs/2009.14153, 2020.

S. Hershey, D. P. W. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal. The benefit of temporally-strong labels in audio event classification. *arXiv preprint arXiv:2105.07031*, 2021.

R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023.

V. Iashin and E. Rahtu. Taming visually guided sound generation. In *BMVC*, 2021.

J. Kong, J. Kim, and J. Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

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*, 2023.

H. Liao, H. Han, K. Yang, T. Du, R. Yang, Z. Xu, Q. Xu, J. Liu, J. Lu, and X. Li. Baton: Aligning text-to-audio model with human preference feedback. *ArXiv*, abs/2402.00744, 2024.

H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. P. Mandic, W. Wang, and M. . Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning*, 2023.

S. Luo, C. Yan, C. Hu, and H. Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models, 2023.

W. S. Peebles and S. Xie. Scalable diffusion models with transformers. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, 2022.

C. Raffel, N. M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2019.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015.

R. Sheffer and Y. Adi. I hear your true colors: Image guided audio generation. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2022.

A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

Z. Xie, X. Xu, Z. Wu, and M. Wu. Picoaudio: Enabling precise timestamp and frequency controllability of audio events in text-to-audio generation. *ArXiv*, abs/2407.02869, 2024.

M. Xu, C. Li, Y. Ren, R. Chen, Y. Gu, W. Liang, and D. Yu. Video-to-audio generation with hidden alignment, 2024.

D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.