Content-Style Disentangled Audio Style Transfer via Diffusion Model

Anonymous ICME submission

***Abstract*—Deep generative models have advanced the synthesis of high-quality audio signals, shifting the focus from audio fidelity to user-specific customization. Despite significant progress, current models struggle to generate style-consistent audio. Audio style transfer offers a more intuitive approach for capturing user intent but faces challenges in the disentanglement and interpretation of content and style. This paper introduces a novel framework for content-style disentangled audio style transfer. We introduce an interpretable, formula-based style distance that effectively disentangles content and style within the language-audio feature space. The proposed QwenAudio-Contrastive Language Audio Pretraining (Qwen-CLAP) content extraction module and the CLAP-based style disentanglement loss coordinated with the style reconstruction loss, enable interpretable disentanglement and stylization. Comprehensive experiments on our new dataset, BBCreatures, demonstrate superior stylization quality, preserving fine style details and original content.**

***Index Terms*—audio style transfer, information disentanglement, latent diffusion model (LDM)**

I. Introduction

Deep generative models have markedly improved the synthesis of high-quality audio, shifting their development focus from mere audio fidelity to user-specific customization [1]. Models like Text-To-Audio (TTA) and Video-To-Audio (VTA), which are based on tuning-free diffusion techniques, show remarkable promise in audio personalization and customization [2]–[5]. However, these models often struggle with producing stylistically consistent outputs, frequently requiring complex prompt engineering which complicates usability [1]. In this context, audio style transfer becomes not just a useful tool, but a necessary one for refining the output of generative models, as it allows for precise stylistic adjustments that align with user preferences without the need for cumbersome prompt tuning [6], [7]. It can serve as an effective post-processing module following the generation of an audio clip by TTA or VTA models, ensuring the final output meets the desired stylistic criteria and offering a more seamless and intuitive experience for users [8].

Given a style reference audio, our goal is to transfer its style to an arbitrary content reference audio. The generated stylization audio should have the same style with the style reference audio, and the same content with the content reference audio. For example, given the sheep bleating sound as content reference audio, and the cat meowing sound as style reference audio, our goal is to generate the same cat meowing sound following the sheep's bleating rhythm as shown in Figure 1. Here we define the sound characteristics of the cat as style and the sound events of the sheep's bleating as content. [1]

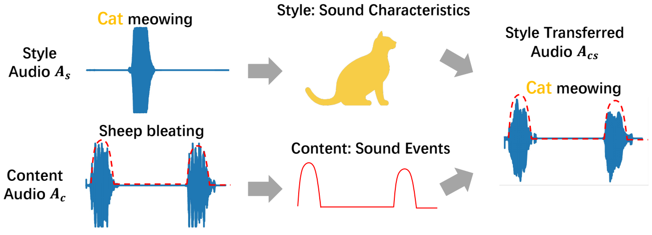

Fig. 1. Concept illustration of audio style transfer.

To fulfill the task of audio style transfer, valuable efforts have been devoted. For instance, methods like CycleGAN-VC [9] and StarGAN-VC [10] have attempted to apply similar principles to audio, where the transfer process is modeled as a separate optimization of content loss and style loss. This approach has achieved impressive results in generating audio that mimics certain styles while preserving the original content, thus inspiring numerous successors.

However, diving deeper into the essence of audio style transfer reveals two fundamental challenges with the existing approaches: I) Overlooking Disentanglement: Content and style information in audio are inherently intertwined, yet they represent distinct aspects of audio signal [11]. When content and style information are not properly disentangled, the resulting transfer can become muddled, with the boundaries between content and style blurring [7]. Without proper disentanglement, the stylization process can introduce artifacts or unwanted distortions, resulting in less-than-ideal transfers where the content is either overly stylized or the style is inadequately applied [2]. II) Lack of Interpretability: The internal mechanisms of deep learning models used for audio style transfer are often opaque, leading to a significant challenge in understanding how the model distinguishes between "content" and "style" [11]. These models typically operate as complete black boxes, where it is difficult to decipher the specific features or representations that the model associates with each aspect of the audio [12]. This lack of interpretability hinders the ability to precisely control the stylization process, as users are unable to understand or adjust how content and style are being manipulated at different stages [6].

To address the challenges of audio style transfer, we propose a novel content and style (C-S) disentanglement framework. The first step is to identify a backbone generation model conditioned on sound events, ensuring that the input audio's content is preserved. Specifically, we use the T-Foley diffusion

[1] Listen to more examples in supplementary material.

model [13] for its superior performance in generating audio that is well-synchronized with temporal events. We then fine-tune it using our proposed methodology to effectively learn disentangled style information while preserving the original content. We then decompose the task of audio style transfer into two distinct subtasks: I) disentangling content and style, and II) transferring style. To disentangle content and style, we leverage text modality, as content is typically easier to describe with text compared to style. This allows us to bypass the ambiguity associated with disentangling style by explicitly extracting content information while implicitly learning the complementary style. Specifically, we introduce the QwenAudio-Contrastive Language Audio Pretraining (Qwen-CLAP) [14], [15] content extraction module to extract domain-aligned content information. To effectively transfer style, we define style information as a style distance in the language-audio-aligned feature space. We also design specific losses to comprehensively capture style information and ensure accurate one-to-one mappings during the fine-tuning process. In particular, we propose the CLAP-based style disentanglement loss, supplemented with the style reconstruction loss, to facilitate precise and consistent style transfer.

Our comprehensive evaluations, including comparisons and ablation studies on our newly proposed BBCreatures dataset, demonstrate the effectiveness and superior performance of our framework. The dataset comprises high-quality audio clips that exhibit clear stylistic differences, ideal for testing our model. With the well-disentangled C-S, our framework achieves very promising stylizations with fine-grained style details, well-preserved contents, and a deep understanding of the relationship between C-S.

The main contributions of the paper are as follows:

- To the best of our knowledge, we are the first to propose an interpretable, formula-based style distance that effectively disentangles content and style within the language-audio-aligned feature space.
- We introduce a new CLAP-based style disentanglement loss coordinated with a style reconstruction loss that facilitates precise style transfer.
- We propose a novel fine-tuned framework based on information disentanglement for audio style transfer, which enhances disentanglement interpretability and output quality in audio style transfer.

II. Related Works

A. Audio Style Transfer Without Disentanglement

Diffusion models have emerged as powerful tools for generating high-quality audio signals [2], [5], [16]. However, few of them have taken C-S disentanglement into account, leading to unrealistic audio transfers due to either excessive or insufficient stylization [1].

Models like AudioLDM [2] and AUDIT [16] rely on textual prompts to guide the generation of audio and require users to provide detailed descriptions of desired styles, which can be inflexible since style can hardly be expressed clearly with text. Similarly, AP-Adapter [3] uses text prompts for music editing tasks but suffers from similar limitations. These models cannot perform style transfer in a flexible and user-friendly manner. Other models like [17] and [18] in the music domain adopt example-based audio style transfer, but do not take explicit C-S disentanglement into account, limiting their performance and flexibility since the content will be either overly stylized or the style will be inadequately applied [12], [19].

B. Audio Style Transfer Using Disentanglement

Disentangling content from style in audio has been explored in various contexts, each addressing different aspects of the challenge but often lacking in interpretability.

Music Mixing Style Transfer uses contrastive learning to separate audio effects from content [6], allowing for style transfer within music mixing. Zero Shot Audio to Audio Emotion Transfer With Speaker Disentanglement [8] aims to transfer emotions between audio clips while maintaining speaker identity. Despite this, the interpretability of what constitutes "style" versus "content" in these works remains an issue. SpeechSplit [12] and AutoVC [19] both are designed around speaker disentanglement, where SpeechSplit further divides speech into rhythm, content, pitch, and timbre components. AutoPST [11] focuses on rhythm disentanglement, separating rhythmic patterns from the rest of the speech content. However, they suffer from a lack of interpretability due to their complex, often opaque mechanisms. This limitation hinders users' ability to understand and control the disentanglement process effectively.

In summary, existing audio style transfer methods either fail to achieve proper C-S disentanglement or lack interpretability in the disentanglement process. To overcome these limitations, our work seeks to propose a novel framework that addresses these fundamental challenges, aiming for more interpretable and higher-quality stylizations.

III. Methodology

Task formulation. Given the style audio A_s and the content audio A_c, our objective is to disentangle their respective style and content, and then transfer the style of A_s to the content of A_c, resulting in the stylized audio A_{cs}. This process aims to achieve a precise fusion where the content of A_c retains its integrity while adopting the stylistic information of A_s.

Method Overview. Figure 2 shows the overview of our model, which consists of three key components: I) the T-Foley style transfer module, II) the Qwen-CLAP content extraction module, and III) the CLAP-based style disentanglement loss coordinated with the style reconstruction loss.

Section III-A introduces the T-Foley diffusion model, detailing its role in our framework. In Section III-B, we elaborate on the process of extracting content information using the QwenAudio Large Language Model (LLM) and the CLAP encoder. Finally, Section III-C delves into the specifics of designing disentanglement loss and reconstruction loss to influence the generation process.

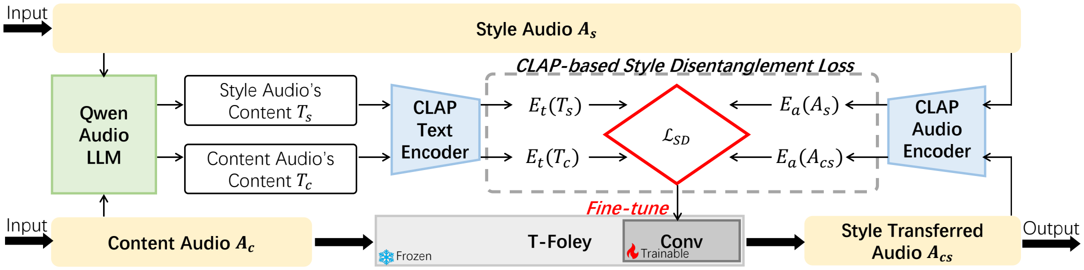

Fig. 2. The overall architecture of the proposed model. Initially, the content audio A_c and the style audio A_s are processed by QwenAudio Large Language Model (LLM) to extract explicit content descriptions, T_c and T_s. Subsequently, A_c is fed into the T-Foley diffusion model to generate the stylized result A_{cs}. During fine-tuning, T-Foley is guided by the CLAP-based style disentanglement loss $\mathcal{L}_{SD}$ together with the style reconstruction loss $\mathcal{L}_{SR}$ (see details in Section III-C2). In this phase, most of the original model's parameters remain frozen, with only the final convolutional layer of T-Foley being trainable.

A. T-Foley Style Transfer Module

T-Foley is a temporal-event-guided waveform generation model for foley sound synthesis which achieves superior performance in synchronization with the temporal events. It generates audio using two conditions: the sound class tag and explicit temporal event feature, i.e., the frame-level envelope features. The training goal of the latent diffusion model (LDM) ϵ_θ is to reconstruct the audio prior z according to the corresponding class embedding and temporal event feature as

$$\mathcal{L}_{LDM} = E_{x,\epsilon \sim N(0,I)} \left|\left|\epsilon - \epsilon_\theta \left(z_\sigma, \sigma, c, T\right)\right|\right|_2^2, \tag{1}$$

where σ is the current denoising time step, c is class embedding, and T is temporal event feature. The core objective of T-Foley is to generate audio that accurately reflects given temporal events. To achieve this, T-Foley uses the Root Mean Square (RMS) value of the waveform as a frame-level envelope feature. For the i-th frame, the RMS is computed as:

$$E_i(x) = \sqrt{\frac{1}{W}\sum_{t=ih}^{ih+W} x^2(t)}, \tag{2}$$

where $x(t)$ is the audio waveform, W is the window length, and h is the hop size. The Event-L1 Distance (E-L1) metric is proposed to evaluate the effectiveness of temporal conditioning. It measures how well the generated sound matches the given temporal event condition by calculating the L1 distance between the RMS features of the target and generated samples:

$$E\text{-}L1 = \frac{1}{k}\sum_{i=1}^{k} ||E_i - \hat{E}_i||, \tag{3}$$

where E_i is the ground truth event feature for the i-th frame, and $\hat{E}_i$ is the predicted event feature.

In our pipeline, the content audio A_c is first processed through the pre-trained T-Foley model ϵ_θ. We then apply the CLAP-based style disentanglement loss along with the style reconstruction loss to fine-tune the reverse diffusion process of the model ($\epsilon_\theta \rightarrow \epsilon_{\hat{\theta}}$), resulting in a stylized output A_{cs} influenced by the style audio A_s. During this fine-tuning phase, we freeze most of the original model's parameters, enabling only the final convolutional layer of ϵ_θ to be trainable. Once fine-tuning is completed, the model can generate a stylized version of any input content audio, capturing the stylistic features of A_s while preserving the original content.

Due to the limited set of sound classes in the released pre-trained model, we retrain T-Foley using our BBCreatures dataset, removing the class condition to expand the range of application. We maintain the temporal condition to ensure control over the output in the temporal domain (see Section IV-A for more details).

In summary, T-Foley provides a robust foundation for our style transfer pipeline by accurately synthesizing sound event guided audio, and its flexibility in fine-tuning allows for the seamless integration of stylistic features while maintaining content consistency in the generated outputs.

B. Qwen-CLAP Content Extraction Module

Instead of employing complex strategies for disentangling content and style from audio, we propose an interpretable and straightforward approach to achieving similar capability. The content extraction module aims at explicitly extracting content information and implicitly learning complementary style information. Compared with the under-determination of style, content is usually easier to describe with text [20]. This textual representation serves as an effective and interpretable proxy for the audio's content information, allowing us to bypass the ambiguity often associated with disentangling style. By transforming the challenge of C-S disentanglement into the task of precise content extraction, our approach simplifies the disentanglement process and enhances interpretability.

To achieve this, we leverage QwenAudio, a state-of-the-art model proven to excel in sound understanding tasks such as Automatic Audio Captioning (AAC) and Audio Question Answering (AQA) [14], alongside CLAP, which captures rich semantic information linking language and audio [15]. Given the style audio A_s and content audio A_c, we employ QwenAudio to generate descriptive sentences T_s and T_c that capture their respective contents, like "There are 3 times of dog barking", "The baby cried twice" etc. These sentences are then encoded using the CLAP text encoder to obtain their corresponding feature representations $E_t(T_s)$ and $E_t(T_c)$. At the same time, we use CLAP audio encoder to extract the features of style audio A_s and content audio A_c. By CLAP's design, where text and audio features share a common feature space, these representations can be manipulated algebraically within this unified feature space. We found that a subtraction operation of audio features and content text features can effectively disentangle content and style information, eliminating the content part of the audio features in an interpretable

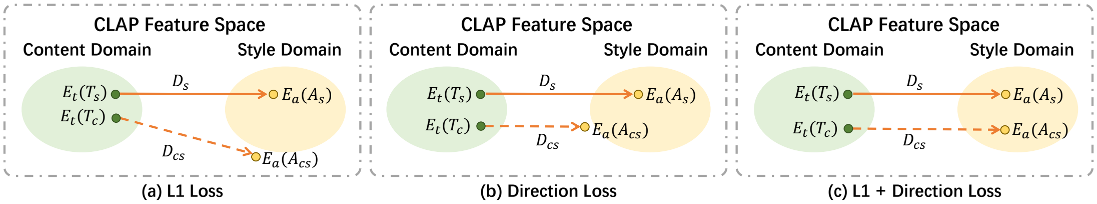

Fig. 3. Illustration of different loss functions to transfer the disentangled style information. (a) L1 loss cannot guarantee the stylized results are within the style domain. (b) Direction loss aligns the disentangled directions but cannot realize accurate mappings. (c) Combining L1 loss and direction can achieve accurate one-to-one mappings from the content domain to the style domain.

way. Through this method, we achieve an interpretable and effective means of disentangling style and content, enhancing the capability for style transfer while maintaining content fidelity.

In summary, our approach capitalizes on the strengths of QwenAudio and CLAP to provide an interpretable mechanism for content extraction and style disentanglement. This not only simplifies the process but also enhances the quality of style transfer by ensuring that only the relevant style information is transferred to the target content.

C. Loss Function and Fine-tuning

1) CLAP-Based Style Disentanglement Loss: After obtaining T_c and T_s—the content descriptions for the content audio A_c and the style audio A_s, respectively—the subsequent step involves learning the disentangled style information of A_s that is complementary to its content. To achieve this, the selected encoder must possess the capability to effectively distinguish between different audio characteristics, particularly in terms of stylistic features [21]. Leveraging previous research [2], [4], [5], which demonstrated CLAP's proficiency in capturing not only semantic information but also being sensitive to stylistic differences, we employ the CLAP text encoder E_t and audio encoder E_a to formulate the disentanglement in a latent semantic space:

$$D_s = E_a\left(A_s\right) - E_t\left(T_s\right), \tag{4}$$

where D_s is the style information vector of style audio A_s in CLAP feature space. The subtraction of content text features from the audio features facilitates the disentanglement of style and content, allowing the CLAP feature space to serve as a metric for measuring "style distance" between content and stylized results. This "style distance" can be interpreted as the disentangled style information. Similarly, we can subtract the text features of content audio A_c from the audio features of the stylization result A_{cs} since A_c and A_{cs} should share the same content information:

$$D_{cs} = E_a\left(A_{cs}\right) - E_t\left(T_c\right), \tag{5}$$

where D_{cs} is the disentangled style information vector of A_{cs}. After obtaining D_s and D_{cs}, the challenge shifts to properly aligning the two. A possible solution is optimizing the L1 loss:

$$\mathcal{L}_{SD}^{L1} = ||\ D_{cs} - D_s||\,. \tag{6}$$

However, as illustrated in Figure 3(a) and further validated in Section IV-B3 Table III, minimizing the L1 loss does not guarantee the stylized result A_{cs} is within the style domain of the style audio A_s. Since L1 loss only minimizes the absolute difference (i.e. Manhattan distance), it can produce stylized audio that satisfies the Manhattan distance but deviates from the target style domain in the transfer direction. To address this issue, we introduce a directional constraint:

$$\mathcal{L}_{SD}^{dir} = 1 - \frac{D_{cs} D_s}{||D_{cs}||\,||D_s||}\,. \tag{7}$$

This direction loss ensures alignment between the style vectors of the original style audio and the stylized results, improving upon the limitations of L1 loss, as illustrated in Figure 3 (b). Combining both losses yields precise mappings from content to style domains, as illustrated in Figure 3 (c). Finally, our style disentanglement loss is defined as a compound of $\mathcal{L}_{SD}^{L1}$ and $\mathcal{L}_{SD}^{dir}$:

$$\mathcal{L}_{SD} = \lambda_{L1}\mathcal{L}_{SD}^{L1} + \lambda_{dir}\mathcal{L}_{SD}^{dir}, \tag{8}$$

where λ_{L1} and λ_{dir} are hyper-parameters set to 10 and 1 in our experiments. Since our style distance is formulated from the difference between the content and its stylized output, we can gain a deeper understanding of the C-S relationship through learning. This approach enhances the interpretability of C-S and C-S disentanglement, resulting in more natural and harmonious style transfers.

2) Style Reconstruction Loss: To fully utilize the information provided by the style audio and enhance the stylization effects, we incorporate a style reconstruction loss at the beginning of the fine-tuning process. Given the style audio A_s, the diffusion model should aim to reconstruct the original style audio as accurately as possible. We define the style reconstruction loss as follows:

$$\mathcal{L}_{SR} = ||\ A_{ss} - A_s\ ||_2^2, \tag{9}$$

where A_{ss} represents the stylized result when A_s also serves as content audio. We optimize this loss separately before optimizing the style disentanglement loss $\mathcal{L}_{SD}$.

In conclusion, the combination of the CLAP-based style disentanglement loss and style reconstruction loss enables effective and interpretable disentanglement and transfer of style information, facilitating more accurate and coherent style transfer. (Refer to Section IV-B3 Table III for validation.)

IV. Experiments

A. Experiments Setup

1) Dataset: Existing audio-text datasets often fail to meet the specific requirement for audio style transfer: each audio clip should contain only one type of sound to ensure that style information pertains solely to a single category. To address this, we constructed a new dataset named BBCreatures from

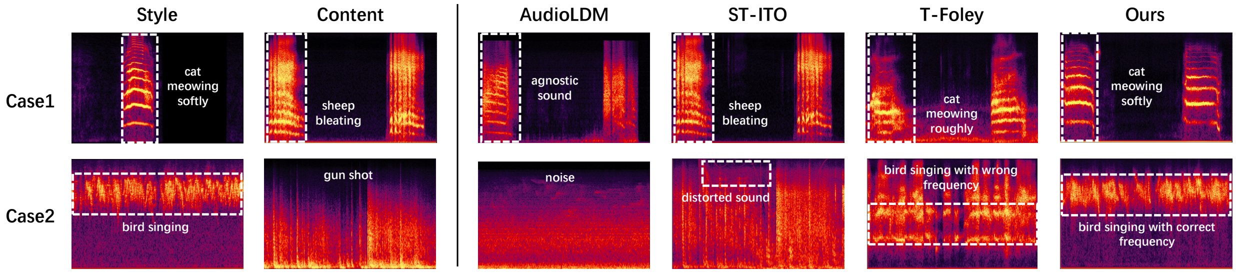

Fig. 4. We visualize the results of various audio style transfer models using mel-spectrograms. On the left, we present the original style and content audio, while the right side displays the corresponding stylized results. Our model outperforms the others by better-preserving content and accurately capturing style. In contrast, the other models either produced poorly transferred audio with low quality or failed to capture the style details effectively.

BBCSFX[2]. BBCSFX comprises 33,066 clips across 1,791 categories. We chose BBCSFX because it contains high-quality audio clips with minimal noise and predominantly single-category sounds. For our style transfer method, we selected five categories—baby, bird, cat, cow, and sheep—that exhibit clear stylistic differences within each category. All audio clips were processed to a sampling rate of 22,050 Hz and cut into 4-second segments. After manually removing silent clips, we prepared 800 clips per category, totaling 4,000 clips. It consists of 3,500 training samples and 500 testing samples, comparable in size to the dataset used for T-Foley [22].

2) Implementation Details: We retrained the backbone diffusion model T-Foley on a single Tesla V100 GPU using the BBCreatures dataset for 500 epochs with a batch size of 8 samples, omitting the class condition while maintaining the same experimental settings as T-Foley. Subsequently, we utilized the pre-trained CLAP encoder from AudioLDM [2] and fine-tuned the model for each category using one style audio and 100 content audios from the testing dataset over 20 epochs at a learning rate of 1×10^{-4}. To improve the efficiency of the fine-tuning process, we reduced the number of denoising steps to 10 during fine-tuning and restored it to 100 during evaluation. Finally, we selected the fine-tuned model with the lowest $\mathcal{L}_{SD}$ loss for each category.

B. Comparative Experiments

1) Qualitative Results: We conducted comparative experiments with recently published state-of-the-art(SOTA) style transfer approaches, including AudioLDM and ST-ITO [7], along with our baseline model T-Foley. Qualitative evaluation focused on two aspects: content alignment with the content audio and style similarity with the style audio. As shown in Figure 4, our model outperformed others in terms of preserving content and capturing style accurately. Other models either produced poorly transferred or low-quality audio (ST-ITO) or generated irrelevant sounds (AudioLDM). While T-Foley was able to transfer some stylistic features, it did not preserve details as effectively as our approach. Additional examples are provided in the supplementary material.

2) Quantitative Results: For objective evaluation, we employed CLAP score, and Resemblyzer[3] score to quantify the style similarity between the generated audio and the style audio. Additionally, we used the Frechet Audio Distance (FAD)[4], and Kullback-Leibler Divergence (KL)[4] to evaluate the content similarity between the generated audio and the content audio. Also, the E-L1 loss proposed by T-Foley is used to prove our fine-tuning process does not corrupt the original content preservation ability of the baseline model.

TABLE I
RESULTS OF OBJECTIVE EVALUATION.

Model	CLAP↑	Resemblyzer↑	E-L1↓	FAD↓	KL↓
AudioLDM	0.68	0.69	0.059	16.8	1.87
ST-ITO	0.48	0.58	0.041	**12.5**	**0.51**
T-Foley (Baseline)	0.71	0.70	0.038	17.6	0.96
Ours	**0.80**	**0.76**	**0.029**	16.6	0.87

Our model outperformed all others on the CLAP, Resemblyzer, and E-L1 metrics while achieving second-best performance on the FAD and KL metrics. ST-ITO, which performed best on FAD and KL, performed worst on the other metrics. This indicates that our approach offers a more balanced solution in terms of overall performance including style similarity and content fidelity using a much smaller dataset.

It is important to note that FAD, KL, and E-L1 metrics only calculate the distribution distance between the generated audio and the content audio but do not consider the distribution distance concerning the style audio. On the other hand, CLAP and Resemblyzer only compute the similarity between the generated audio and the style audio without considering the content audio. Therefore, these metrics cannot comprehensively evaluate model performance as a whole, making subjective evaluation essential.

For subjective evaluation, we obtain the Mean Opinion Score (MOS) in three main aspects: the realism of the generated audio (MOS-R), content consistency with the content audio (MOS-C), and style similarity with the style audio (MOS-S), as shown in Table II. We collected data from 20 participants who rated 20 examples from five categories. Our model outperformed the baseline and SOTA models in all aspects. Detailed MOS results are provided in the supplementary material.

3) Ablation Studies: To verify the effectiveness of each loss term used for fine-tuning, we present the results of ablation studies in Table III. The full model ($\mathcal{L}_{SD}^{L1}+\mathcal{L}_{SD}^{dir}+\mathcal{L}_{SR}$) outperformed all other configurations on the CLAP and Resemblyzer scores, indicating superior style transfer capabilities. It also

[2] https://sound-effects.bbcrewind.co.uk/

[3] https://github.com/resemble-ai/Resemblyzer

[4] https://github.com/haoheliu/audioldm_eval

TABLE II
Results of subjective evaluation.

Model	MOS-R↑	MOS-C↑	MOS-S↑
AudioLDM	2.50	3.36	1.97
ST-ITO	2.64	3.64	1.65
T-Foley(Baseline)	2.55	3.48	2.86
Ours	**3.65**	**4.04**	**4.05**

achieved the lowest FAD, KL, and E-L1 loss, highlighting its effectiveness in preserving content. The addition of the style reconstruction loss played a crucial role in improving both style transfer and content preservation.

To further verify the necessity of C-S disentanglement, we also experimented without content extraction, using only the CLAP audio encoder. This modification, which bypasses C-S disentanglement, resulted in poor performance, even after adding the style reconstruction loss. This reinforces the importance of disentangling C-S for optimal performance.[5]

TABLE III
Ablation Studies on loss function. *** denotes our full model.**

Loss	CLAP↑	Resemblyzer↑	E-L1↓	FAD↓	KL↓
$\mathcal{L}_{SD}^{L1}$	0.75	0.69	0.081	19.8	1.72
$\mathcal{L}_{SD}^{L1} + \mathcal{L}_{SD}^{dir}$	0.79	0.68	0.080	18.6	1.55
$\mathcal{L}_{SR}$	0.73	0.67	0.041	21.6	4.11
$\mathcal{L}_{\mathbf{SD}}^{\mathbf{L1}} + \mathcal{L}_{\mathbf{SD}}^{\mathbf{dir}} + \mathcal{L}_{\mathbf{SR}}*$	**0.80**	**0.76**	**0.029**	**16.6**	**0.87**
$\mathcal{L}_{CLAP\text{-}audio}$	0.73	0.71	0.034	17.5	1.33
$\mathcal{L}_{CLAP\text{-}audio} + \mathcal{L}_{SR}$	0.76	0.74	0.030	16.9	0.96

V. Conclusion

In this work, we addressed the challenge of generating style-consistent audio by proposing a novel framework for content-style disentangled audio style transfer using diffusion models. We introduced an interpretable, formula-based style distance that effectively disentangles content and style within the language-audio-aligned feature space. Through the integration of the Qwen-CLAP content extraction module and the CLAP-based style disentanglement loss coupled with a style reconstruction loss, we demonstrated the effectiveness and interpretability of our approach through extensive experimental evaluations on the newly developed BBCreatures dataset.

References

[1] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li, "A survey on generative diffusion models," *IEEE Transactions on Knowledge and Data Engineering*, 2024. I, II-A

[2] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023. I, I, II-A, II-A, III-C1, IV-A2

[3] Fang-Duo Tsai, Shih-Lun Wu, Haven Kim, Bo-Yu Chen, Hao-Chung Cheng, and Yi-Hsuan Yang, "Audio prompt adapter: Unleashing music editing abilities for text-to-music with lightweight finetuning," *arXiv preprint arXiv:2407.16564*, 2024. I, II-A

[4] Junwon Lee, Jaekwon Im, Dabin Kim, and Juhan Nam, "Video-foley: Two-stage video-to-sound generation via temporal event condition for foley sound," *arXiv preprint arXiv:2408.11915*, 2024. I, III-C1

[5] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen, "Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds," *arXiv preprint arXiv:2407.01494*, 2024. I, II-A, III-C1

[6] Junghyun Koo, Marco A. Martinez-Ramirez, Wei-Hsiang Liao, Stefan Uhlich, Kyogu Lee, and Yuki Mitsufuji, "Music mixing style transfer: A contrastive learning approach to disentangle audio effects," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2022. I, I, II-B

[7] Christian J Steinmetz, Shubhr Singh, Marco Comunità, Ilias Ibnyahya, Shanxin Yuan, Emmanouil Benetos, and Joshua D Reiss, "St-ito: Controlling audio effects for style transfer with inference-time optimization," *arXiv preprint arXiv:2410.21233*, 2024. I, I, IV-B1

[8] Soumya Dutta and Sriram Ganapathy, "Zero shot audio to audio emotion transfer with speaker disentanglement," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10371–10375. I, II-B

[9] Takuhiro Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 2100–2104, 2018. I

[10] H. Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks," *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 266–273, 2018. I

[11] Kaizhi Qian, Yang Zhang, Shiyu Chang, Jinjun Xiong, Chuang Gan, David Cox, and Mark A. Hasegawa-Johnson, "Global rhythm style transfer without text transcriptions," *ArXiv*, vol. abs/2106.08519, 2021. I, II-B

[12] Kaizhi Qian, Yang Zhang, Shiyu Chang, David Cox, and Mark A. Hasegawa-Johnson, "Unsupervised speech decomposition via triple information bottleneck," in *International Conference on Machine Learning*, 2020. I, II-A, II-B

[13] Yoonjin Chung, Junwon Lee, and Juhan Nam, "T-foley: A controllable waveform-domain diffusion model for temporal-event-guided foley sound synthesis," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 6820–6824. I

[14] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023. I, III-B

[15] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "Clap learning audio concepts from natural language supervision," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. I, III-B

[16] Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al., "Audit: Audio editing by following instructions with latent diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 71340–71357, 2023. II-A, II-A

[17] Hong Huang, Yuyi Wang, Luyao Li, and Jun Lin, "Music style transfer with diffusion model," *ArXiv*, vol. abs/2404.14771, 2024. II-A

[18] Sifei Li, Yuxin Zhang, Fan Tang, Chongyang Ma, Weiming Dong, and Changsheng Xu, "Music style transfer with time-varying inversion of diffusion models," *ArXiv*, vol. abs/2402.13763, 2024. II-A

[19] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark A. Hasegawa-Johnson, "Zero-shot voice style transfer with only autoencoder loss," *ArXiv*, vol. abs/1905.05879, 2019. II-A, II-B

[20] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen, "Instantstyle: Free lunch towards style-preserving in text-to-image generation," *arXiv preprint arXiv:2404.02733*, 2024. III-B

[21] Zhizhong Wang, Lei Zhao, and Wei Xing, "Stylediffusion: Controllable disentangled style transfer via diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7677–7689. III-C1

[22] Keunwoo Choi, Jae-Yeol Im, Laurie M. Heller, Brian McFee, Keisuke Imoto, Yuki Okamoto, Mathieu Lagrange, and Shinosuke Takamichi, "Foley sound synthesis at the dcase 2023 challenge," *ArXiv*, vol. abs/2304.12521, 2023. IV-A1

[5] Listen to examples in supplementary material.