Customized Condition Controllable Generation for Video Soundtrack

Supplementary Material

6. Supplementary Material

Our supplementary materials are divided into three parts: the experimental code, the generated sample audio, and the textual materials. The textual materials consist of five sections: Sec. 6.1 introduces the structure and training details of our diffusion model. Sec. 6.2 explains the steps and training details of the Video-Sound-Music Pretraining. Sec. 6.3 provides a detailed analysis of the mechanism of Spectrum Divergence Masked Attention, as proposed in Sec. 3.2, from the perspective of the spectrogram. Sec. 6.4 showcases the performance of the music and sound effects generated by our method in the spectrogram. Sec. 6.5 compares the results of four custom condition generation methods with the pre-optimization results.

6.1. Video-to-Soundtrack Diffusion in Details

Data Sample Processing: For all audio signal processing, we convert all music samples to the required format for training using a 16 kHz sampling rate. For video signal processing, we use a frame rate of 10 fps, following the standards of Film Score and HIMV. Each input pair is a 10.24-second clip randomly selected from the dataset. Short-Time Fourier Transform (STFT) and Mel-spectrogram calculations are performed using a hop size of 160, window size of 1024, filter length of 1024, and 128 Mel-frequency bins.

Model Architecture: Our foundational framework is modified based on AudioLDM. AudioLDM UNet consists of 4 encoder blocks, 1 bottleneck block, and 4 decoder blocks, where each block contains 2 residual CNN layers and 1 spatial Transformer layer. The channel sizes in the encoder blocks are 128, 256, 384, and 640, and the decoder blocks mirror these channel sizes in reverse order. Unlike the AudioLDM U-Net, we replace the standard cross-attention module in each Transformer layer with the Spectrum Divergence Masked Attention (SDMA) module. To ensure feature dimension alignment, the cross-attention module in each SDMA layer has the same shape as the corresponding cross-attention module in the Transformer layer.

Training Details: For training the UNet, we use a batch size of 32 and the AdamW optimizer with a base learning rate of 3×10^{-5}. During the forward process, we use a linear noise schedule from $\beta_1 = 0.0015$ to $\beta_{1000} = 0.0195$ over 1000 steps. During training, we apply classifier-free guidance with a 10% probability and a guidance scale of $w = 2.5$. During the sampling process, we use the DDIM sampler with 200 steps. The latent diffusion model is trained on 4 NVIDIA A800 GPUs. For the VAE and HiFiGAN models, we use the official open-source version of MusicLDM, with their parameters kept frozen during the UNet training.

Implementation of Comparison Model: For baselines of AudioLDM-VA and Tango-VA, we use the official model of AudioLDM and Tango, only modifying the text condition to a video condition. The training details remain consistent with those described above.

6.2. Video-Sound-Music Pretraining Details

For video-sound-music pretraining hyperparameters, we refer to the official repository to conduct the training process of CLAP. We use fully connected layers to map each 1D vector of video, music, and sound effects into a 512-dimensional space, simultaneously training them to align. During training, we use a batch size of 256 and the Adam optimizer with hyperparameters $\beta_1 = 0.99$ and $\beta_2 = 0.9$, along with a warm-up strategy and cosine learning rate decay, with a base learning rate set to 1×10^{-4}.

6.3. Explaining the Effectiveness of Spectrum Divergence Masked Attention

In this study, the spectrograms of the generated audio and the original audio from the dataset appear in Fig. 4 and Fig. 5 respectively. Fig. 5 provides more detailed descriptions of the music and sound effects. The sound effects and music components shown in the audio samples in Fig. 5 are separated from the original audio samples using a source separation model, and then converted into spectrograms using Short-Time Fourier Transform (STFT).

Overall, the spectrogram of the audio sample does not simply equate to the sum of the spectrogram contours of its sound effects and music components. The audio sample's spectrogram contains additional detailed information, some of which overlaps in the primary time and frequency domains. Therefore, the black and red boxes are used merely to highlight the main regions corresponding to the respective concepts and should not be interpreted as implying that the music component, represented by the red box, is completely devoid of the texture corresponding to the sound effect, represented by the black box. The reverse is also true.

From a more detailed perspective, the background noise of the sound effects in the spectrogram is generally much higher than that of the music. This difference is primarily due to the fact that musical tones are typically clearer and more stable than sound effects, resulting in a more concentrated frequency composition and lower background noise in the spectrogram. In contrast, sound effects may contain more irregular components or a broader frequency range, leading to higher background noise. Consequently, the presence of sound effects in the spectrogram can interfere with

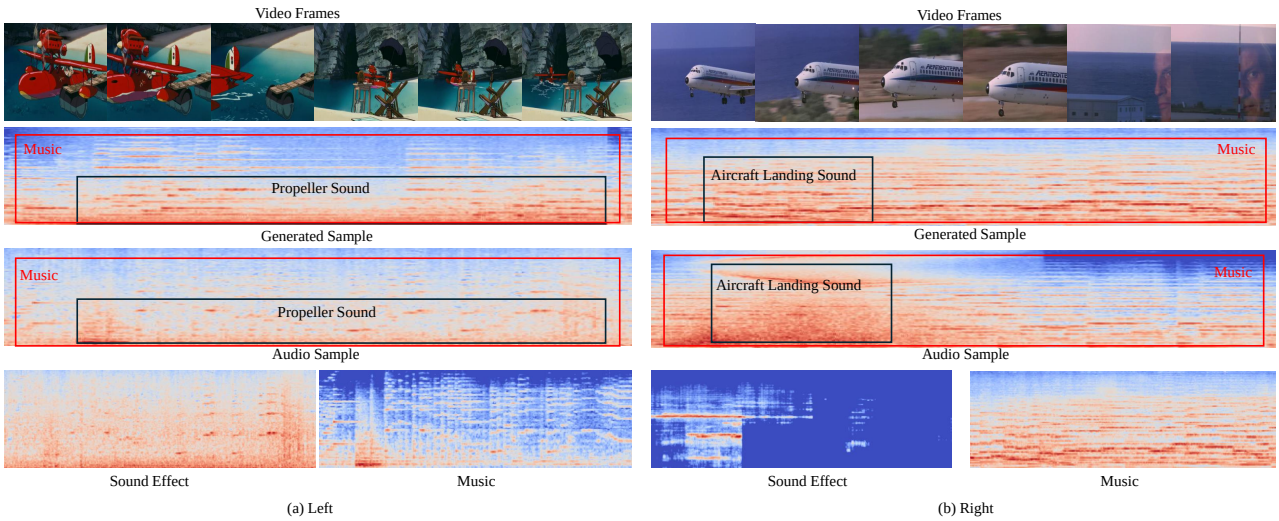

Figure 5. Comparison of mel spectrograms between the audio generated by the generative model and the audio sample from the original dataset. To highlight the difference between music and sound effect, we separate the original audio into Music and Sound Effect.

the music itself, manifesting as the background noise of the sound effects potentially obscuring the texture of the music, making it harder to distinguish. This is why we employ the Spectrum Divergence Masked Attention technique: it aims to mitigate the interference of sound effect noise on the music information, thus enabling a more effective integration of sound effects and music.

Moreover, this indirectly explains why generating texture details using both the music and sound effect control conditions from the video yields far better results than relying solely on the video-generated output. It is one of the key reasons why our diffusion model outperforms both Audioldm-VA and Tango-VA across multiple metrics.

6.4. Explaining the Generation Performance of Sound Effects and Music

From the perspective of generative performance, compared to the original audio, the audio that our model generates exhibits lower background noise, with clearer spectral texture details and more distinct layers. As we show in Fig. 5, the spectrogram of the generated audio demonstrates significantly lower background noise across the entire time domain, particularly in the high-frequency range, compared to the original audio. At the same time, we focus on processing the main sound effects in the video. In terms of spectrogram texture, our model more effectively highlights and distinguishes the musical texture details compared to the original audio. As we show in Fig. 5, the spectrogram reveals that the generated helicopter sound is primarily concentrated in the time domain corresponding to the appearance of the helicopter in the scene, whereas the original audio spreads the helicopter sound effect across the entire video. Moreover, we not only significantly reduce the background noise in the time domain of non-helicopter sound effects but also manage to reduce the background noise during the helicopter sound effect. Furthermore, as we see in the musical section of the spectrogram in Fig. 5, the music texture that our model generates exhibits more pronounced layering.

6.5. Customized Condition Generation

Music-Visual Rhythm Syn Soundtrack Generation: Fig. 6 illustrates the effectiveness of our method in optimizing music content based on the rhythm of video frames. To facilitate comparison, we analyze the spectral details of audio that is generated before and after we apply the Music-Visual Rhythm Syn optimization. For clarity, the video rhythm variation curve is displayed in the Video Rhythm graph, where the horizontal axis represents time, and the vertical axis indicates the magnitude of changes in video frames. Higher vertical values correspond to more dramatic video rhythm variations.

In Fig. 6, yellow boxes highlight regions with significant video rhythm changes. These annotations are reflected in the video frames, the video rhythm curve, the original audio spectrogram, and the optimized audio spectrogram. From the spectrograms, it is evident that our Music-Visual Rhythm Syn optimization method aligns the texture structure of the generated spectrogram with the rhythm variation trend that is shown in the Video Rhythm graph, while maintaining the overall structure of the audio spectrogram. This demonstrates that we successfully adjust the music rhythm to match the video rhythm.

Moreover, our method does not affect the sound effect

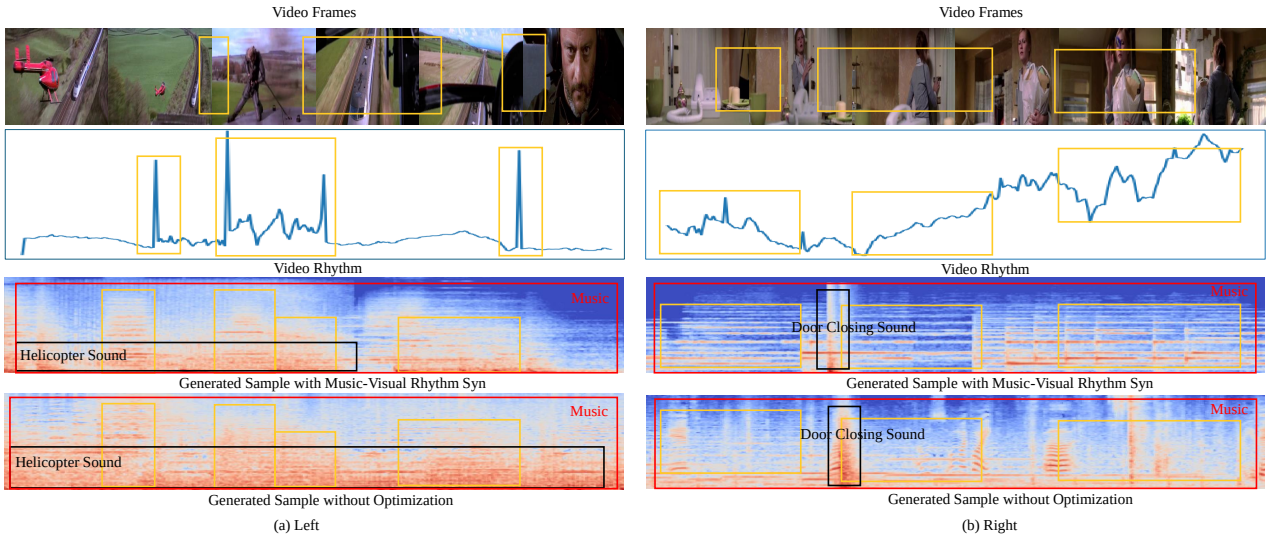

Figure 6. Comparison of mel spectrograms between the audio generated by the music-visual rhythm syn and the generated sample without optimization.

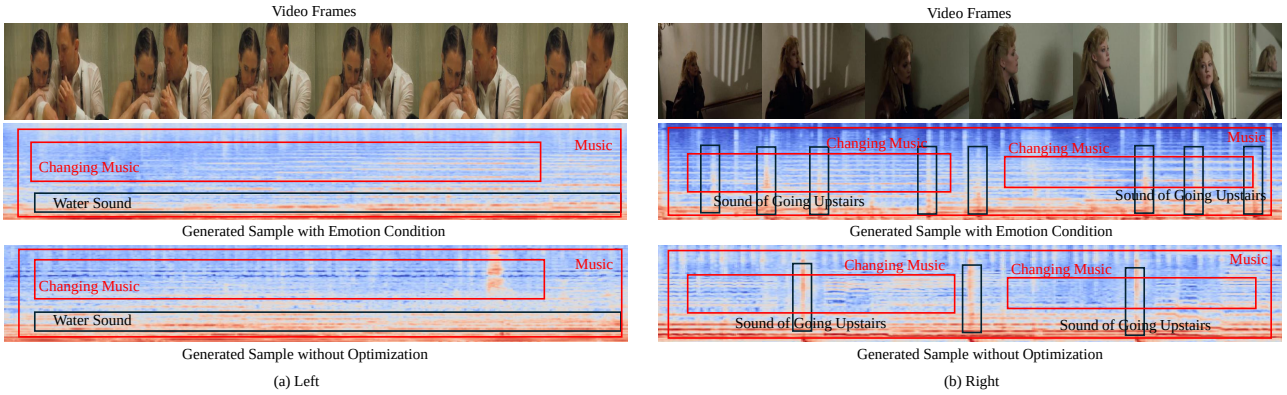

Figure 7. Comparison of mel spectrograms between the audio generated by the emotion condition and the generated sample without optimization.

regions in the spectrogram, preserving the core audio features that are generated by the diffusion model.

Emotion Conditioned Soundtrack Generation: To avoid significant deviation between the optimized audio content and the video content, the text descriptions used for zero-shot classification do not directly include specific video content. Instead, we constrain the descriptions to simple emotional statements, such as "This is a sad piece of music" or "This is a happy piece of music". The results of emotion-conditioned optimization are shown in Fig. 7. The text description for Fig. 7a is "This is a sad piece of music," while for Fig. 7b, it is "This is a tense piece of music." From the spectrograms, we observe that the "sad" spectrogram tends to exhibit smoother textures, whereas the "tense" spectrogram demonstrates more rhythmic textures.

Style Conditioned Soundtrack Generation: Fig. 8 illustrates the results of optimizing audio content based on a specific musical style. To facilitate comparison, we present not only the spectrograms of the audio before and after optimization but also the spectrogram of the reference style music. In the reference style spectrogram, red boxes are used to highlight the textures that represent the musical style. As shown in Fig. 8, we successfully capture and learn the characteristics of the target musical style.

Multi-conditioned Soundtrack Generation: Fig. 9 demonstrates the effect of optimizing music content with multiple custom control conditions. As shown in Fig. 9a, when the number of custom control conditions increases, the optimized audio effects change slightly compared to the pre-optimized version, but overall, they still align with the video content. However, as shown in Fig. 9b, when the region of the spectrogram's audio effects overlaps signifi-

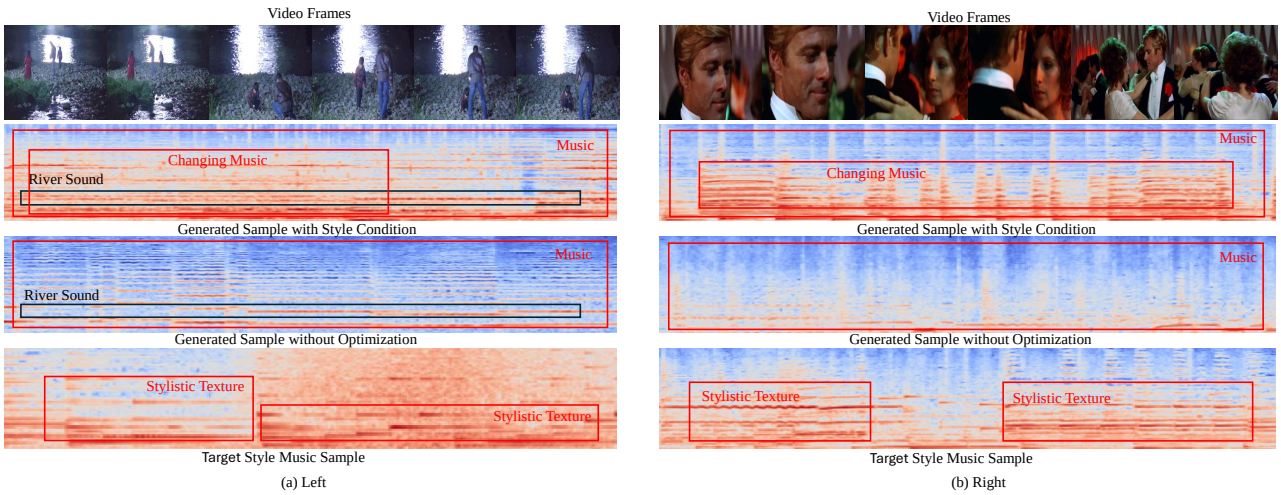

Figure 8. Comparison of mel spectrograms between the audio generated by style condition and the generated sample without optimization.

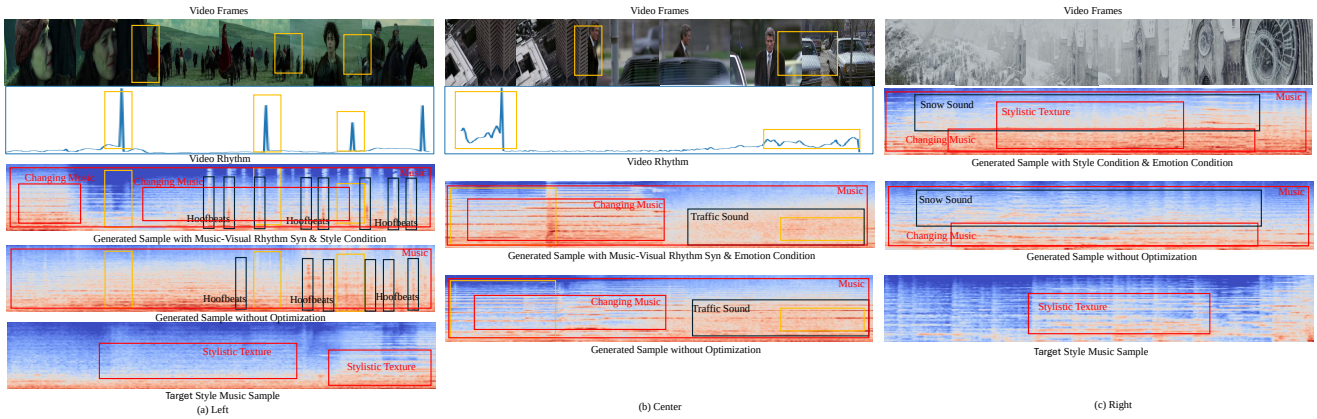

Figure 9. Comparison of mel spectrograms between the audio generated by multi condition and the generated sample without optimization.

cantly with the video's rhythmic changes, this impacts the way the video's rhythm influences the music's rhythm. The emotional text description for Fig. 9b is "This is a heavy piece of music," while for Fig. 9c, it is "This is a cheerful piece of music." From Fig. 9b and Fig. 9c, we can see that under multiple conditions, emotional parameters still effectively control changes in the music's texture. Similarly, Fig. 9a and Fig. 9c show that the style condition still influences the structure of the music's texture.