

183 This appendix provides additional details to complement the main text and support a deeper un-  
184 derstanding of our work. Sec. A investigates how the FAD score varies with changes in loudness.  
185 Sec. B elaborates on the details of Stage 1 in the method described in Sec. 2, including a visual  
186 representation of the generation process. Sec. C supplements the subjective evaluation by offering  
187 further explanation and presenting a representative sample.

## 188 A Loudness-FAD relationship

189 We observed that loudness has a noticeable impact on the Fréchet Audio Distance (FAD)[4], which  
190 evaluates how closely generated audio resembles reference ground truth music in terms of statistical  
191 properties. To investigate this effect, we conducted a dedicated experiment examining how FAD  
192 varies with changes in loudness. Specifically, we normalized the loudness of outputs from three  
193 baseline methods and our approach to a range between -24 and -3 dB LUFS, and then calculated the  
194 FAD scores for each loudness level. As shown in Fig.2, the results consistently reveal that the FAD  
195 score reaches its minimum—indicating the best match to real music—when the loudness is between  
196 -18 and -12 dB LUFS. Interestingly, this range coincides with the dominant loudness levels of the  
197 reference ground truth, which likely reflects the most comfortable listening range for the human  
198 ear. This experiment also suggests that while higher loudness can enhance the masking effect of  
199 noise, excessively high levels can degrade perceptual quality. Therefore, we aim to keep the overall  
200 loudness within an appropriate range to achieve better blending and provide the most comfortable  
201 experience for the listener.

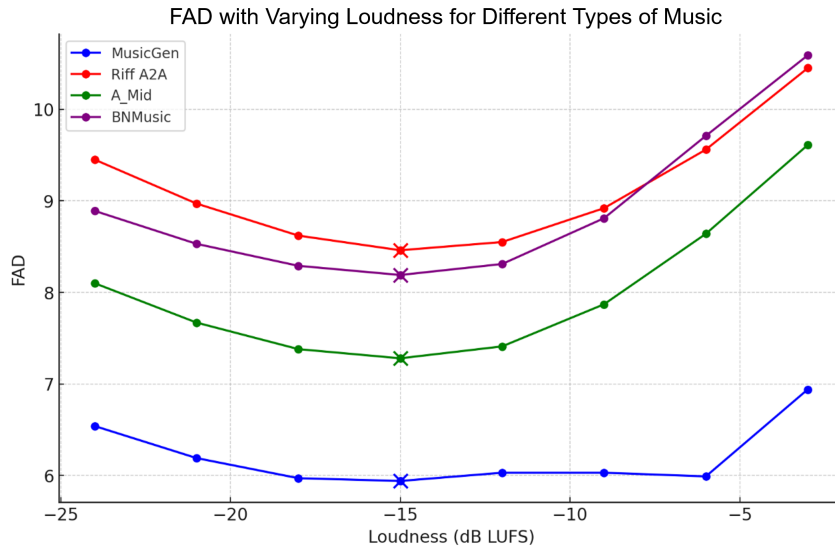


Figure 2: The lowest FAD values for each type of music are highlighted, and they all appear in the -15 dB LUFS, indicating that for all types of music in our experiment, an optimal loudness level consistently falls between -18 to -12 dB LUFS.

## 202 B More details in Stage 1

203 This section provides a detailed supplement to Stage 1 of the method described in Sec. 2. Through  
204 visualizations, we illustrate the motivation for enforcing rhythmical alignment and explain how our  
205 Stage 1 achieves this alignment via a combination of outpainting and inpainting. This approach not  
206 only ensures temporal coherence but also preserves sufficient musicality in the generated content.  
207 Additionally, we describe the strategy used to select the core area of the noise mel-spectrogram plot,  
208 which serves as the foundation for the alignment process.

209 **The significance of alignment.** Our approach leverages generative models to produce music  
210 that rhythmically aligns with the background noise. This alignment facilitates natural blending, as  
211 illustrated in Fig. 3, reducing potential conflicts between the two sources. By ensuring temporal

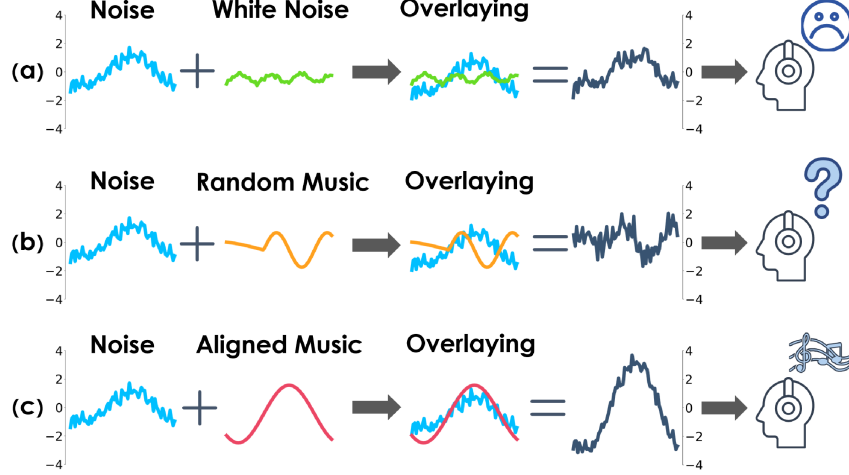


Figure 3: Illustration of how conventional auditory masking, often using overly low volume and mismatched rhythm, can disrupt the listening experience. Proper volume adjustment and rhythmic alignment are essential for achieving a more harmonious and pleasant blend with background noise.

212 coherence from the outset, the combined audio avoids introducing additional disturbances into the  
 213 soundscape. As a result, the subsequent adaptive amplification can be applied more conservatively  
 214 while still achieving effective masking. Even in frequency regions where complete masking is not  
 215 possible, the improved alignment enhances perceptual harmony and helps minimize the listener’s  
 216 awareness of the underlying noise.

217 **The significance of all steps.** Fig. 4 illustrates the detailed process of transforming a single noise  
 218 sample, represented as  $\mathbf{x}_{\text{Noise}}$ , into the final output music in an image representation,  $\mathbf{x}_{\text{Music}}$ . As shown  
 219 in Fig. 4, the outpainting step primarily focuses on diffusing information from the preserved core  
 220 region of the noise  $\tilde{\mathbf{x}}_{\text{Noise}}$  into the surrounding areas. This diffusion embeds contextual information into  
 221 the surrounding music during the generation of  $\mathbf{x}_{\text{Mid}}$ . However, at this stage, directly converting  $\mathbf{x}_{\text{Mid}}$   
 222 into an audio signal  $A_{\text{Mid}}$  would retain noise content from the core region, significantly degrading the  
 223 listening experience. To address this, a subsequent inpainting step is required to mask the remaining  
 224 core noise area and replace it with structured, harmonious music that aligns with the text prompt.  
 225 During this inpainting process, the information previously embedded in the surrounding music during  
 226 outpainting diffuses back into the core region, ensuring seamless integration. The result,  $\mathbf{x}_{\text{Music}}$ ,  
 227 represents a cohesive and complete musical piece.

228 Furthermore, as demonstrated in Fig. 5, both our approach and the Riffusion’s audio-to-audio  
 229 generation [2] exhibit the most effective alignment with the noise. The results of our approach,  
 230 the Riffusion [2]’s, and the MusicGen [1]’s are all generated conditioned on the noise, expected to  
 231 maintain a strong rhythmic consistency for a more seamlessly blending. In contrast, the result of  
 232 MusicGen’s melody-conditioned generation [1], as well as the randomly chosen music, fail to achieve  
 233 similar rhythmic synchronization with the noise, as expected. This highlights the superior ability  
 234 of our method to align the generated music with noise, making it more coherent and seamlessly  
 235 integrated while maintaining pleasant to the ear.

236 **Visualization-Based Comparison.** To further illustrate the blending behavior, Fig. 6 presents  
 237 several representative examples. Each group consists of five plots: the mel-spectrogram of a noise  
 238 sample, followed by four heatmaps showing the difference between that noise and four types of  
 239 music—Random Music, MusicGen, Riffusion-A2A, and our BNMusic. All music samples are  
 240 loudness-normalized to match the noise before computing the difference, ensuring a fair comparison.  
 241 The heatmaps visualize the energy difference between the music and the noise. Red indicates positive  
 242 differences, blue indicates negative, and darker colors represent larger magnitudes. These maps reveal  
 243 how closely each music sample aligns with the noise in terms of spectral energy distribution. Among  
 244 the four, Random Music shows the largest mismatch with the noise, especially in less active frequency  
 245 bands. MusicGen also differs notably, but to a lesser extent. In contrast, Riffusion-A2A and our  
 246 BNMusic demonstrate much closer alignment to the noise across the frequency spectrum. Their

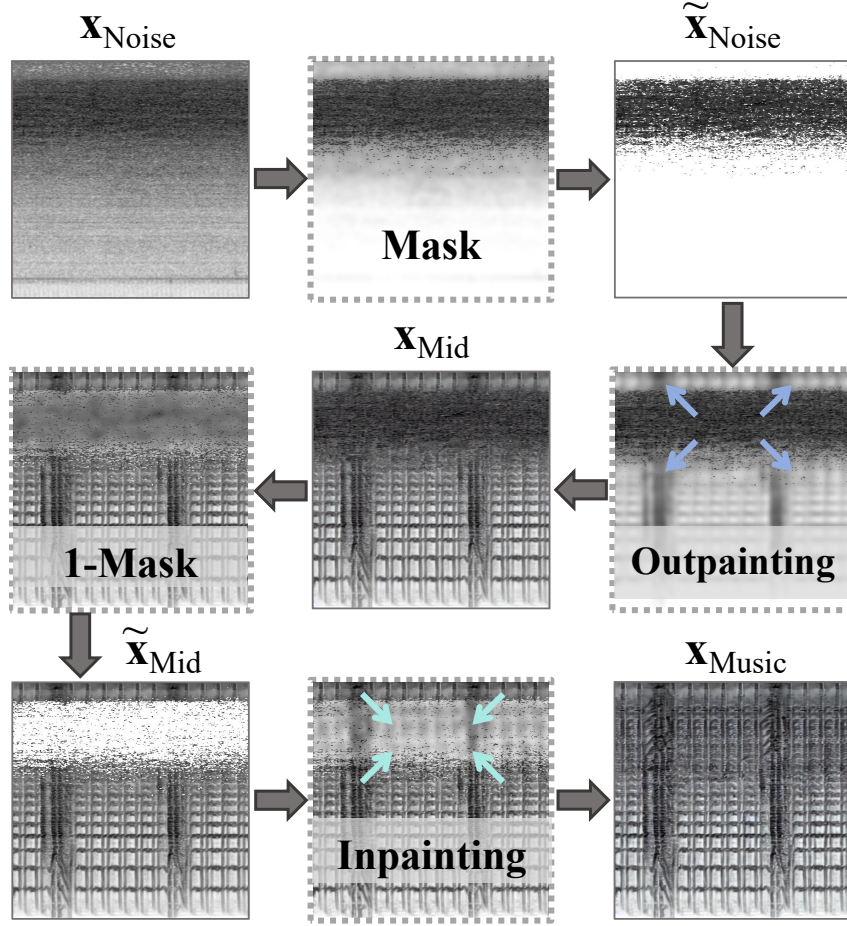


Figure 4: The more detailed illustration for the generation process in Stage 1, transitioning from  $x_{\text{Noise}}$  to  $x_{\text{Music}}$ , should emphasize the distinct processing stages and the regions primarily affected during each step. This would include highlighting how the inpainting step serves as the pivotal transformation within the process, where chaotic noise regions are replaced with structured and meaningful music content.

differences are more evenly distributed and less extreme, indicating better spectral blending. This suggests that A2A and BNMusic are more effective in matching the energy profile of the noise, which may underlie their superior auditory integration. However, both FAD and subjective evaluation results confirm that our BNMusic significantly outperforms A2A in terms of pleasantness and harmony.

**The strategy of picking thresholds.** The selection of 10%–20% of the area with smaller pixel values as the mask region is based on empirical observations. Since the mask is extracted pixel-wise with a value range of 0–255 while the smaller pixel value indicates the higher energy level of the mel-frequency, small variations in pixel intensity can lead to significant differences in the mask area, especially in images with relatively low contrast. Our goal is to ensure that the mask region captures the primary high-energy frequency areas while keeping its size minimal. This approach provides the model with greater flexibility to generate the desired musical elements. Conversely, during the inpainting phase, the preserved core region may sometimes occupy a relatively small proportion of the overall area. In such cases, the limited space can make it challenging for inpainting to generate sufficiently detailed and coherent musical content. To address this, we adjust the threshold to slightly enlarge the mask area, enabling the generation of a more complete, harmonious, and cohesive musical result.

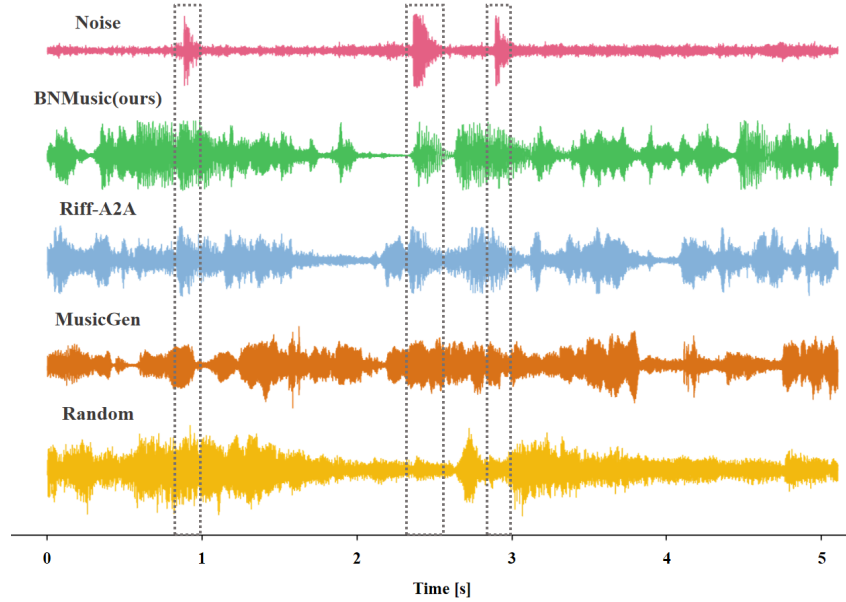


Figure 5: The waveforms of a set of samples, consisting of noise, a random music track, and three music segments generated based on the noise, are shown. As highlighted, our method achieves one of the best alignment effects, where any impulsive sound from the noise is seamlessly blended with a corresponding strong musical sound, ensuring a smooth integration between the two.

## C More details about subjective evaluation

This section provides additional details regarding the subjective evaluation process. It outlines the evaluation protocol and criteria used to assess perceptual quality, and includes a sample of the questionnaire presented to listeners during the study.

The participants would be seeing these words:

OVL (Overall): Measures the overall quality and pleasantness of the audio.

Perceptibility: Indicates how noticeable the original noise is in the presence of the music.

Both metrics are rated on a scale from 1 to 5, where 1 represents the least pleasant sound or the noise being most perceptible, while 5 denotes the most pleasant sound or the noise being least perceptible.

Each participant was presented with a set of audio clips, including the original noise, three music clips generated using the noise, and a randomly selected real music piece all overlaid with the noise. The participants were asked to rate the overall quality and perceptibility of each clip. A sample page of the questionnaire is presented in Fig. 7

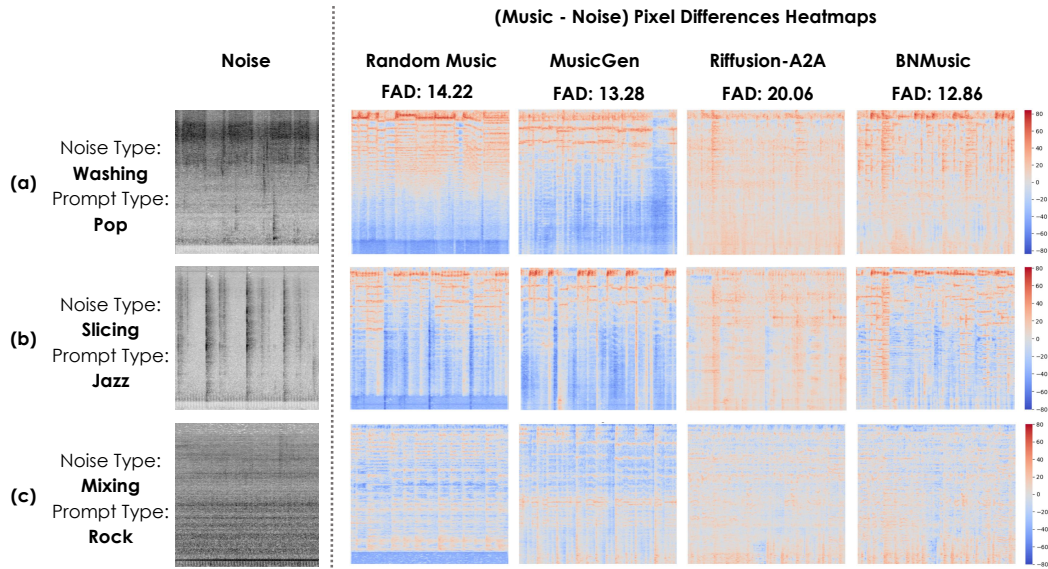


Figure 6: **Visualization of noise-music blending effectiveness across methods.** The left images display the mel-spectrograms of three types of noise, while the right heatmaps show the differences between the generated music and the noise. The heatmaps illustrate how four music samples blend with the respective noise. Red represents positive values, blue indicates negative values and darker colors correspond to larger magnitudes, highlighting the blending effectiveness of each music type.

1

Noise playing simultaneously with 4 different music or audios

Fill in with '1-5' scores,  
1 -> the least pleasant/most perceptible  
5 -> the most pleasant/least perceptible.

Scoring 1-5		a	b	c	d
Original Noise	a				
	b				
	c				
	d				
OVL					
Perceptibility					

OVL (Overall): Measures the overall quality and pleasantness of the audio.

Perceptibility: Indicates how noticeable the original noise is in the presence of the music.

Both scores range from 1 to 5, where 1 denotes the least pleasant or the noise being most perceptible, and 5 denotes the most pleasant sound or the noise being least perceptible.

Figure 7: A sample page of the subjective evaluation.