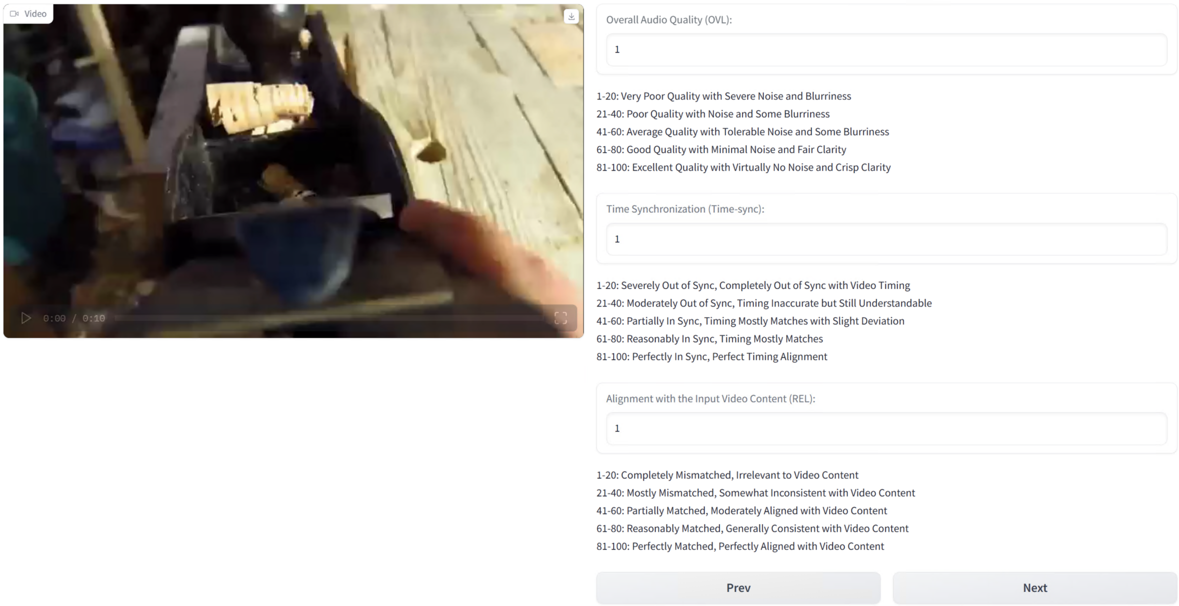
SonicVisionLM: Playing Sound with Vision Language Models

Supplementary Material



Overall Audio Quality (OVL):

1

1-20: Very Poor Quality with Severe Noise and Blurriness
21-40: Poor Quality with Noise and Some Blurriness
41-60: Average Quality with Tolerable Noise and Some Blurriness
61-80: Good Quality with Minimal Noise and Fair Clarity
81-100: Excellent Quality with Virtually No Noise and Crisp Clarity

Time Synchronization (Time-sync):

1

1-20: Severely Out of Sync, Completely Out of Sync with Video Timing
21-40: Moderately Out of Sync, Timing Inaccurate but Still Understandable
41-60: Partially In Sync, Timing Mostly Matches with Slight Deviation
61-80: Reasonably In Sync, Timing Mostly Matches
81-100: Perfectly In Sync, Perfect Timing Alignment

Alignment with the Input Video Content (REL):

1

1-20: Completely Mismatched, Irrelevant to Video Content
21-40: Mostly Mismatched, Somewhat Inconsistent with Video Content
41-60: Partially Matched, Moderately Aligned with Video Content
61-80: Reasonably Matched, Generally Consistent with Video Content
81-100: Perfectly Matched, Perfectly Aligned with Video Content

Figure 1. The evaluation page for the unconditional generation task. We display a screenshot of the primary test interface that participants will encounter. Each participant is required to input three scores for the presented content. Upon clicking the 'Next' button, they will be directed to the subsequent video in the test sequence.

A.1. Subjective Results Details

We provide the screenshot of the main evaluation page the participant will see during the test in Fig.1.

A.2. Timestamp Detection Module Precision Experiment

Our timestamp detection model shows promising performance, with an Average Precision(AP) of 0.80 and an accuracy of 0.72 on the Greatest Hits test set, and even higher results on the validation set with an AP of 0.92 and accuracy of 0.82. However, its performance on the CountixAV test set is comparatively lower, achieving an AP of 0.52 and an accuracy of 0.52. This discrepancy likely stems from the complexity of sound sources in our training dataset, leading to potential inaccuracies in ground truth. Such labelling challenges can adversely affect recognition accuracy, particularly in scenarios involving non-static footage.

A.3. Additional Results

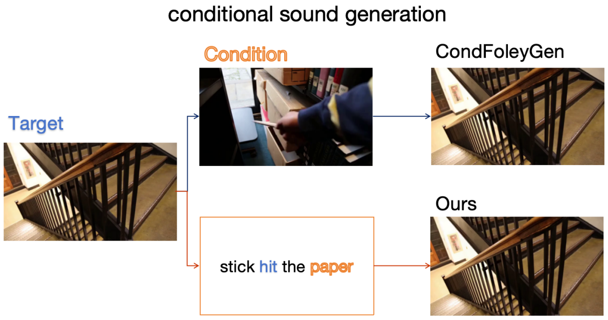


Figure 2. screenshot of the conditional sound generation task section in the demo video.

Conditional Generation Results. As shown in Fig.2, the left column represents the target video, the middle column showcases the control conditions for CondFoleyGen and our model, and the right column displays the generated results. We provide 6 examples for the previously mentioned conditional generation task, and the corresponding audio outcomes can be observed in the demo video.

Unconditional Generation Results. As shown in Fig.3, we compare a comparison of the results generated by GT, SpecVQGAN, DIFF-FOLEY, and our model. We provide 6 examples from CountixAV, and the audio samples are available in the demo video.

Figure 3. screenshot of the unconditional sound generation task section in the demo video.

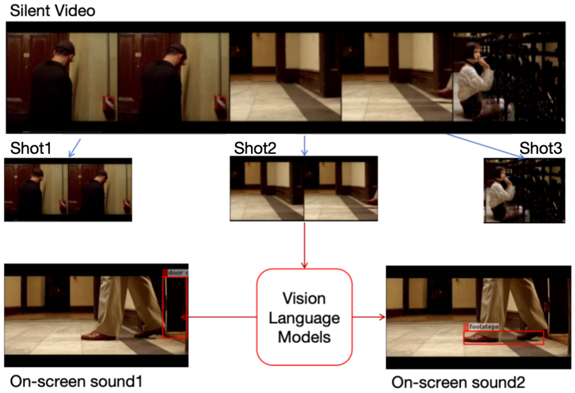


Figure 4. The process of video-to-text and text-based interaction components in multi-track generation tasks.

Multi-soundtracks Generation Results. As shown in Fig.4, we segment the video into different shots first. Shots with clear actions are directly processed by the VLMs to obtain corresponding sound effect descriptions and their spatial positioning within the video. These descriptions are used to generate on-screen sound. Other shots accept user editing and are then fed into the LDM to produce off-screen sounds. We provide a simple and a complex case respectively in the demo video.