

A Demo Page

We’ve prepared a **demo page**, included in the supplementary materials, to illustrate our method and showcase our results. **We strongly encourage you to visit this webpage and experience our results.** For the best viewing experience, we recommend using Google Chrome, as the page may not be fully compatible with Safari.

On the demo page, you’ll find:

- **Interactive Interface Demo:** We’ve built a Gradio interface, making it easy to use our separation method. It supports both video and audio uploads, and allows for selecting different base models, inversion methods, and hyperparameters. The output is the separated audio, enabling you to effortlessly try out our approach.
- **Separation Results from Various Methods:** We present separation results across diverse scenarios, including musical instrument sources, daily events, and more.

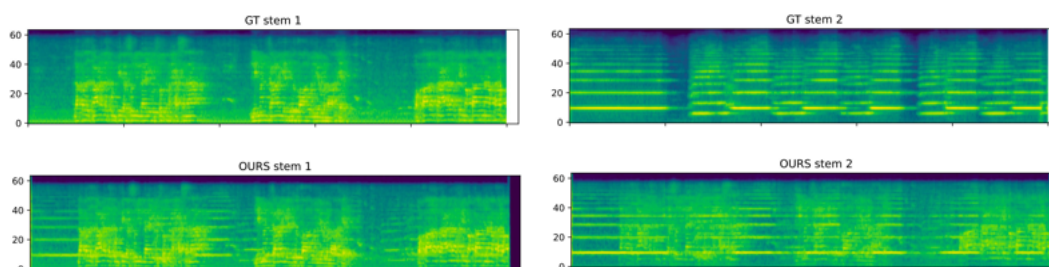


Figure 4: Failure case analysis of ZeroSep. Mixture: Man speech (stem 1) + Shofar (stem 2).

B Failure Case Analysis

While generally effective, ZeroSep can sometimes fail to fully isolate the target source. This typically occurs when an interfering source possesses significant energy that the model cannot eliminate in a single iteration. An illustrative example of such a failure is presented in Fig. 4. We postulate that, given the inherent progressive operation of diffusion models, the removal of interfering sources also proceeds incrementally. Consequently, this performance limitation may be tied to the number of inference steps utilized. Potential avenues for improvement include increasing the inference steps or iteratively applying the separation process.

C More Separation Results

These figures present mel-spectrograms that visualize the audio separation performance on two-source mixtures. For each figure, the rows are ordered from top to bottom as follows: the first source’s Ground Truth, followed by its separation results from LASS-Net, FlowSep, AudioEdit, AudioSep, and Ours. This sequence is then repeated for the second source: Ground Truth 2, LASS-Net 2, FlowSep 2, AudioEdit 2, AudioSep 2, and Ours 2. You might notice some white or empty areas on the right side of the mel-spectrograms; these are simply due to the varying lengths of the audio samples.

MUSIC - cello--Gdh8N_KpLy+erhu-DVyVd_QUCI8

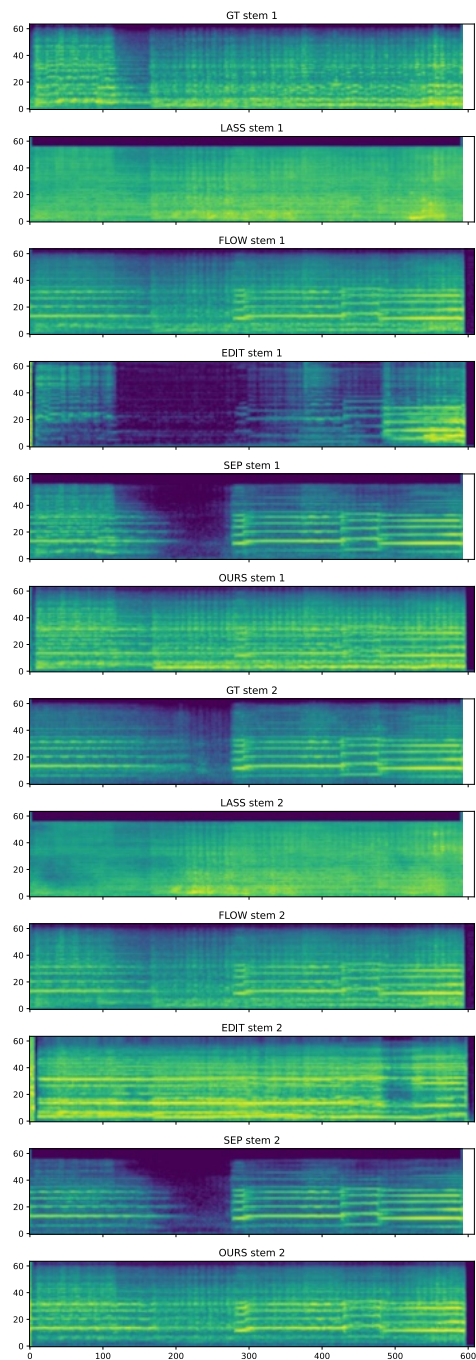


Figure 5: Mixture: Cello (stem 1) + Erhu (Stem 2)

MUSIC - acoustic_guitar-Pzf9MQKkoNM+tuba-4NvJEIaXgo

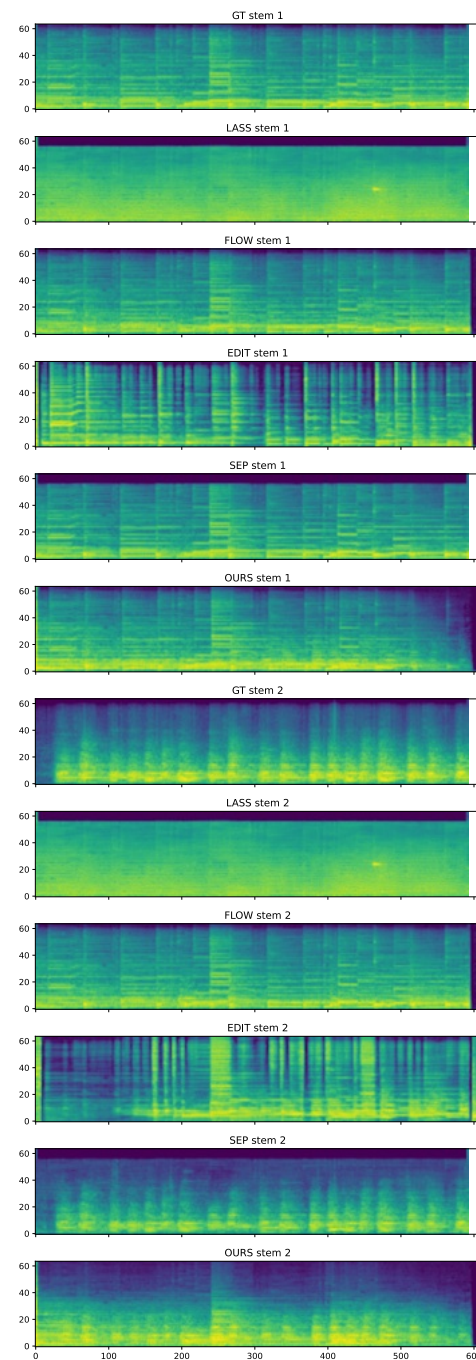


Figure 6: Mixture: Acoustic Guitar (stem 1) + Tuba (Stem 2)

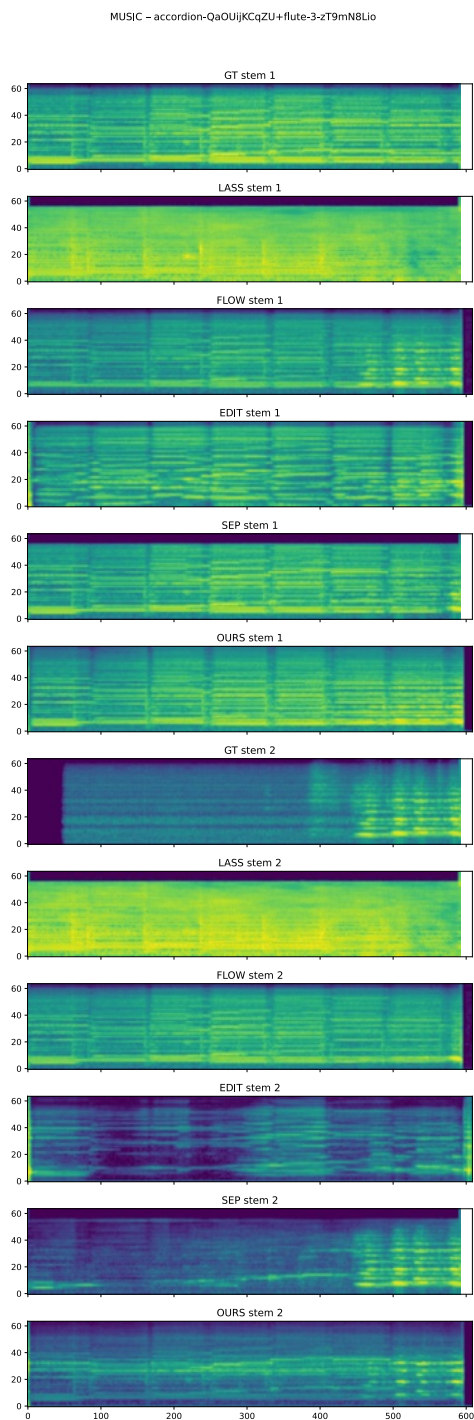


Figure 7: Mixture: Accordion (stem 1) + Flute (Stem 2)

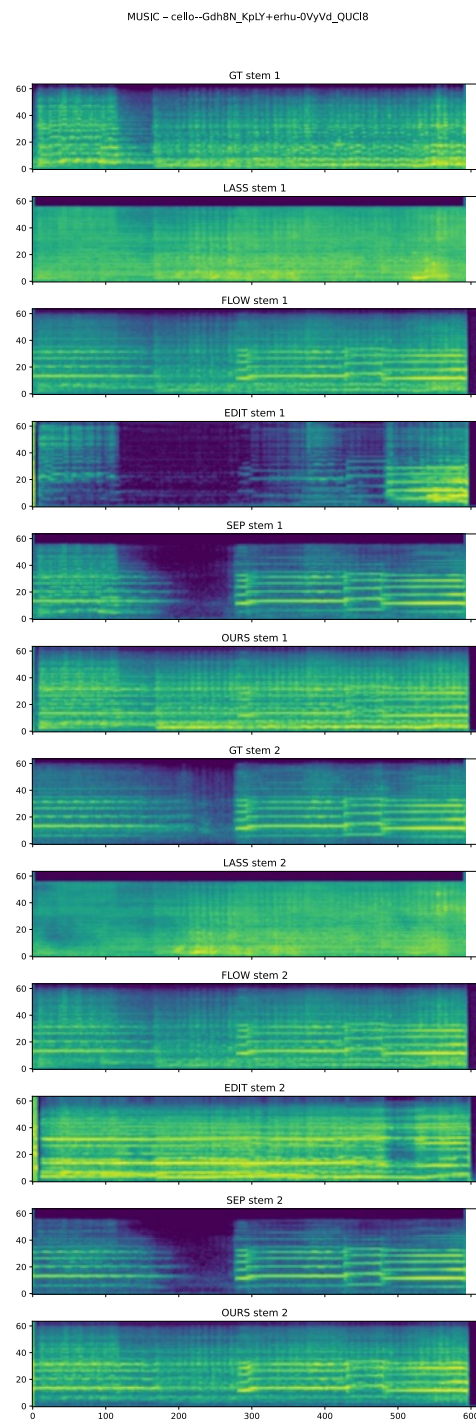


Figure 8: Mixture: Cello (stem 1) + Erhu (Stem 2)

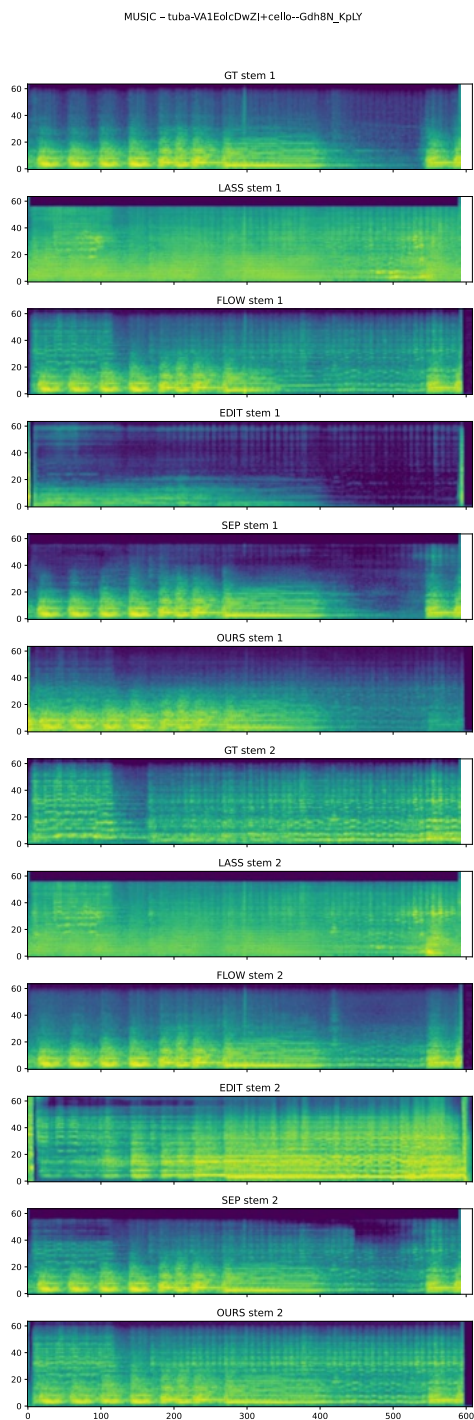


Figure 9: Mixture: Tuba (stem 1) + Cello (Stem 2)

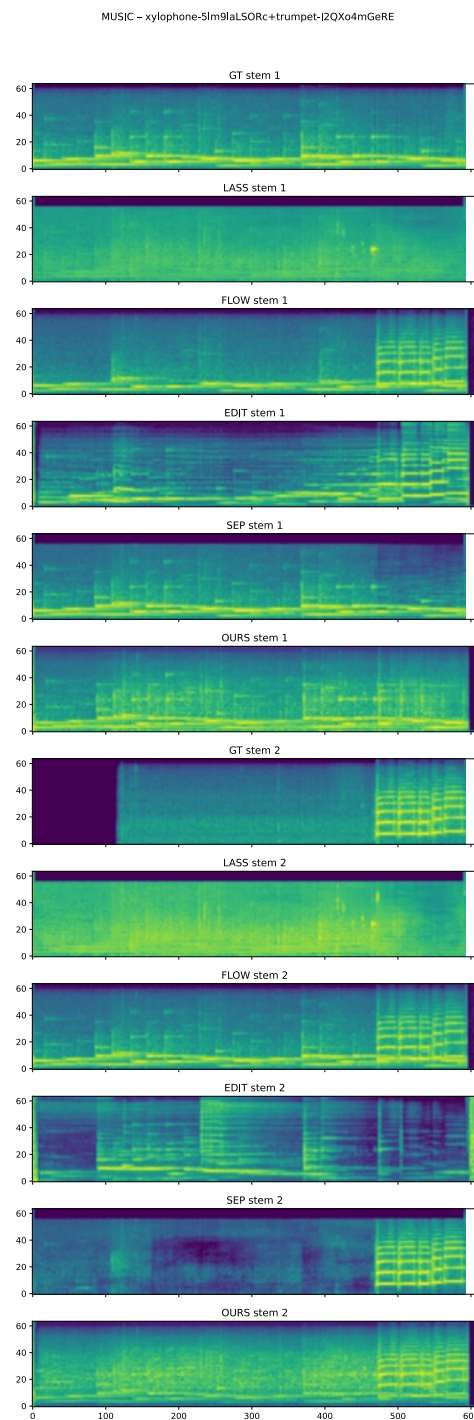


Figure 10: Mixture: Xlyophone (stem 1) + Trumpet (Stem 2)

AVE -- 8jVnlHDsso-Truck+2vj4gKp_sag-Banjo

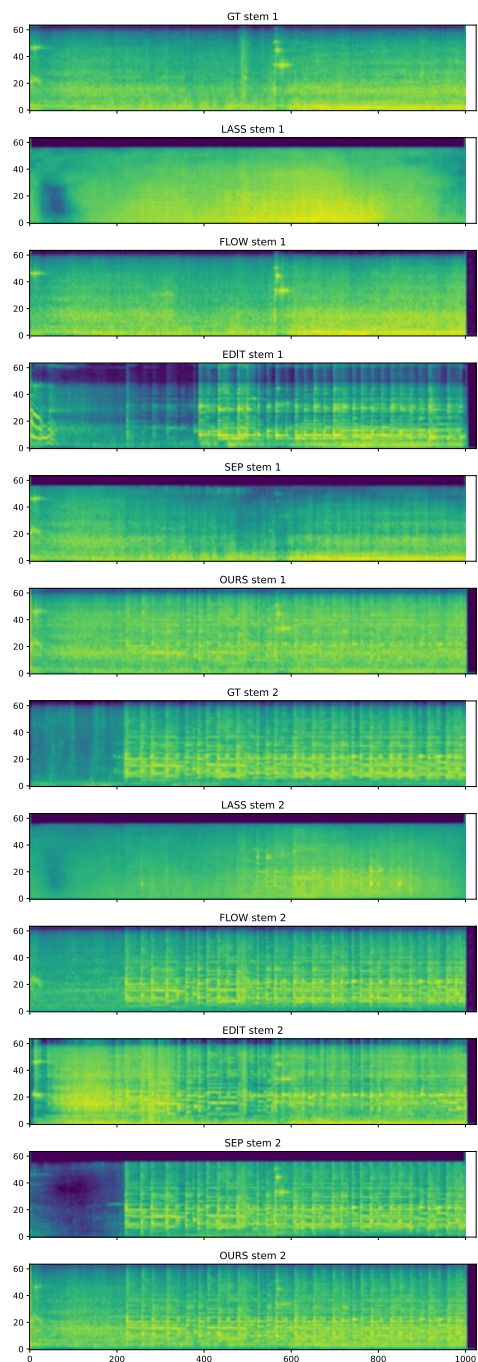


Figure 11: Mixture: Truck (stem 1) + Banjo (Stem 2)

AVE -- 9ummbDsgFM-Chainsaw+1NYCiPBzn-E-Accordion

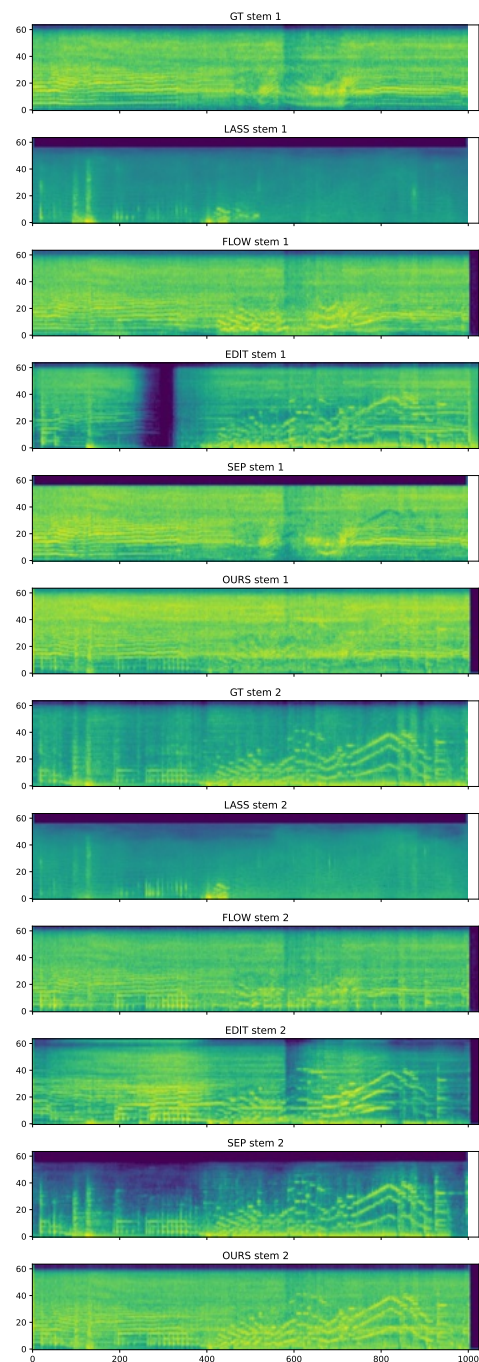


Figure 12: Mixture: Chainsaw (stem 1) + Accordion (Stem 2)

AVE --5QrBL6MzLg-Train horn+FRp2fWka7s-Bark

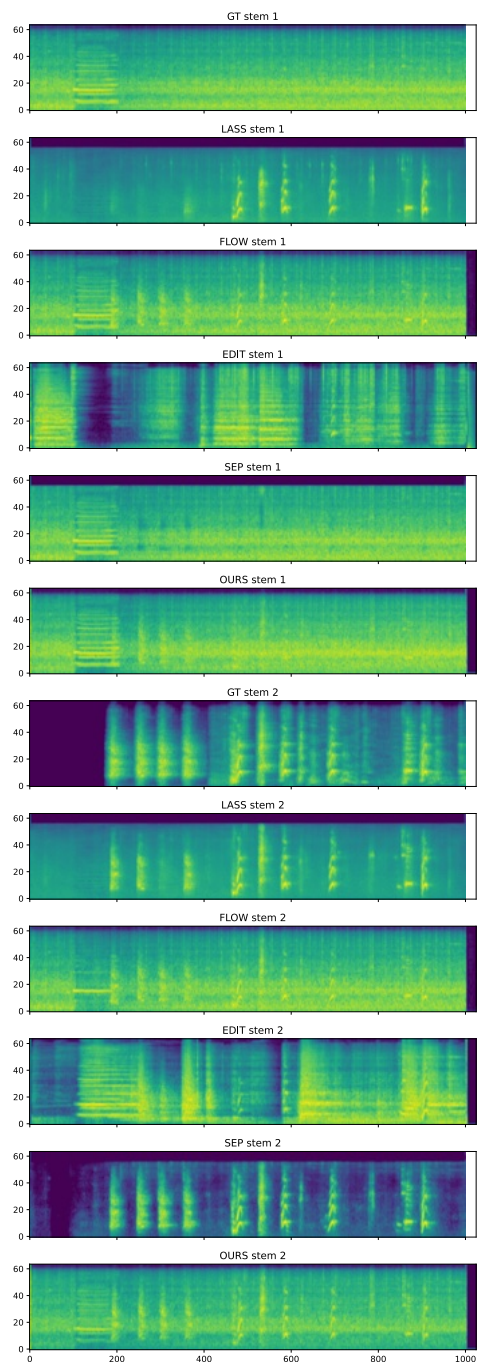


Figure 13: Mixture: Train Horn (stem 1) + Bark (Stem 2)

AVE --en7GAdXAQk-Male speech, man speaking+8B4pp_c9c0E-Fixed-wing aircraft, airplane

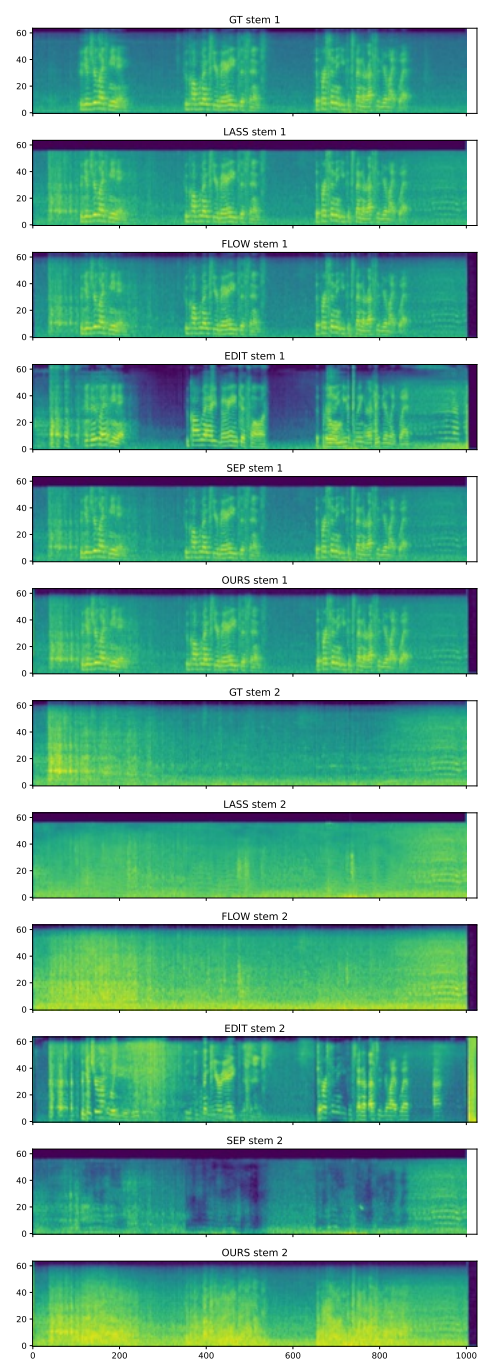


Figure 14: Mixture: Male Speech (stem 1) + Airplane (Stem 2)

AVE -- lKMo9-20Zc-Truck+16cp5o6bBCE-Ukulele

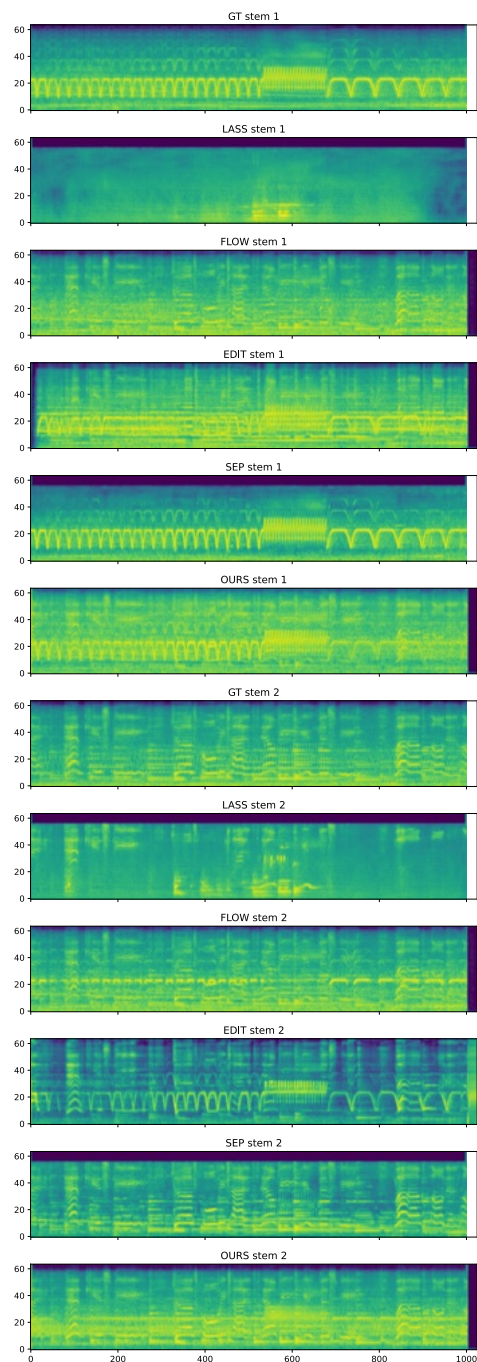


Figure 15: Mixture: Truck (stem 1) + Ukulele (Stem 2)

AVE -- BJNMHMZDcU-Bark+-2C9ZpNhlvg-Toilet flush

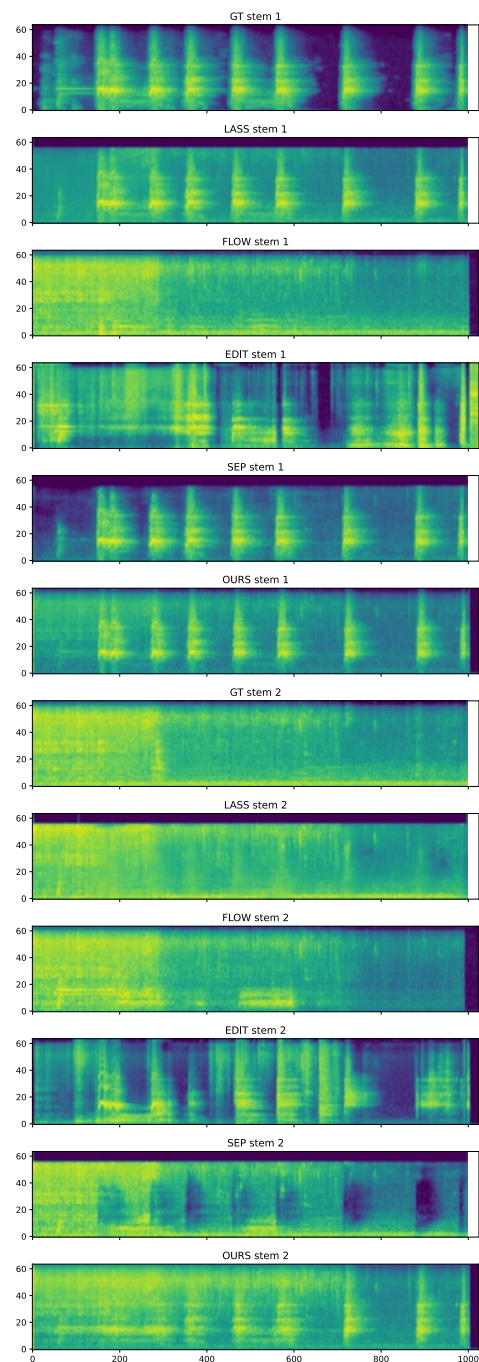


Figure 16: Mixture: Bark (stem 1) + Toilet Flush (Stem 2)