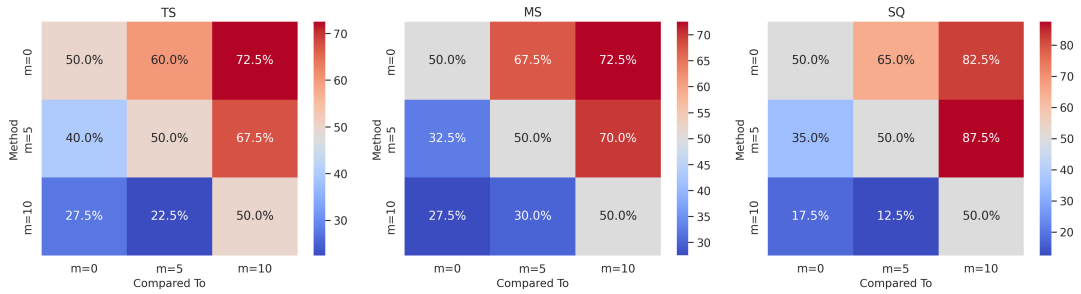


Fig. 1: The win rate heatmaps across TS / MS / SQ (%) with different degree of poor performance.

A. SUBJECTIVE EVALUATION DETAILS

Among the 20 participants, 6 had no prior music education, 9 had limited musical training, 3 had several years of musical experience, and 2 were professionals currently working in the music industry.

For the listening test's pairwise comparison, we provide a win-rate heatmap Figure 1 as a visual representation of the results. Regarding participant feedback, some participants noted that samples with a higher perceived degree of poor performance were primarily characterized by degraded audio quality, rather than by deficiencies in recorder playing technique. This feedback suggests that the attribute vector might mainly influence the perceived cleanliness of timbre, potentially at the expense of fidelity, rather than capturing nuanced aspects of poor performance such as pitch instability or lack of tonguing.

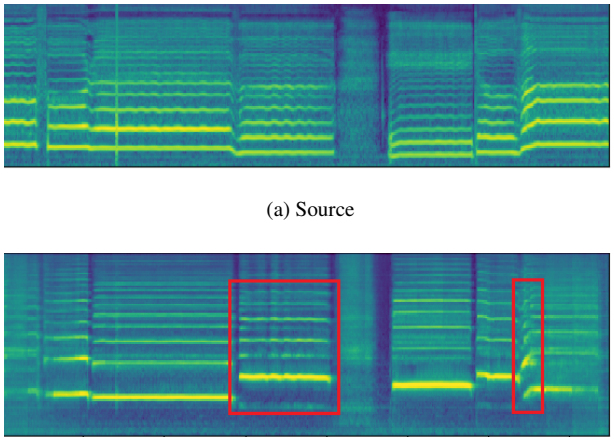
B. ABLATION STUDY

Given the large timbral gap between human vocals and the failed recorder, the latent loss $\mathcal{L}_{\text{Latent}}$, which encourages the encoder to align content representations across domains, may negatively affect conversion quality. To investigate its impact, we compare a VAE-GAN variant that disables both pitch-shifting augmentation and latent loss (w/o Aug. + $\mathcal{L}_{\text{Latent}}$) against the variant without augmentation alone (w/o Aug.).

As shown in Table 1, w/o Aug. + $\mathcal{L}_{\text{Latent}}$ outperforms the variant without augmentation alone (w/o Aug.) across FAD, Pitch Class L_1, and DTW distance. These results suggest that removing latent loss improves pitch accuracy and melody preservation. However, the removal of latent loss also prevents the encoder from filtering out timbre-specific information. As a result, some converted outputs retain vocal-specific timbral characteristics that are nearly infeasible on the failed recorder. Figure 2 illustrates this phenomenon: the red rectangular parts in the Mel spectrogram of the converted audio (Figure 2b) highlight *vibrato* and *glissando* patterns that closely resemble those in the source audio (Figure 2a). While these expressive techniques are natural in human singing, they are difficult to produce on the failed recorder. This observation underscores the critical role of latent loss in suppressing domain-specific timbral traits within the universal encoder.

Table 1: Ablation study on augmentation and $\mathcal{L}_{\text{Latent}}$.

Setting	**FAD** ↓	**Pitch Class** L_1	**DTW Distance** ↓
VAE-GAN w/o Aug.	8.51	3.18	1.99
VAE-GAN w/o Aug. + $\mathcal{L}_{\text{Latent}}$	**7.68**	**1.61**	**1.24**



(a) Source

(b) VAE-GAN w/o Aug. + $\mathcal{L}_{\text{Latent}}$

Fig. 2: The Mel spectrograms of (a) the source vocal audio and (b) the converted output from the VAE-GAN variant without augmentation and latent loss ($\mathcal{L}_{\text{Latent}}$).