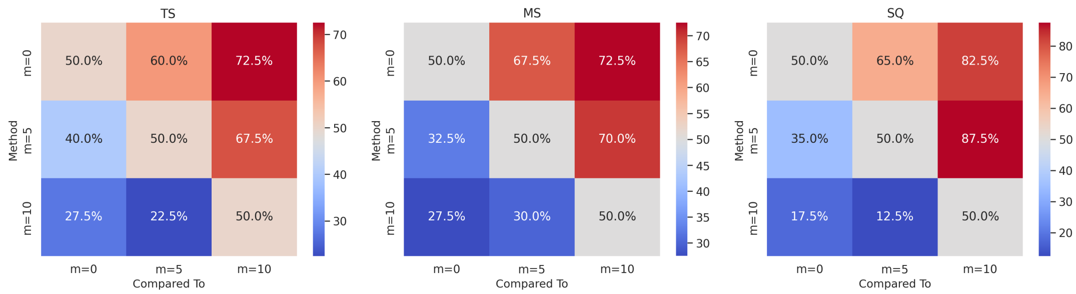


Fig. 1: The win rate heatmaps across TS / MS / SQ (%) with different degree of poor performance.

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. 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 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

B.1. Without Latent Loss

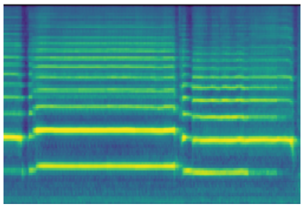
Given the large timbral gap between human vocals and the failed recorder, the latent loss $\mathcal{L}_{\text{Latent}}$, which encourages the encoder to align content representations across domains, may negatively affect conversion quality. To investigate its impact, we compare a VAE-GAN variant that 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.

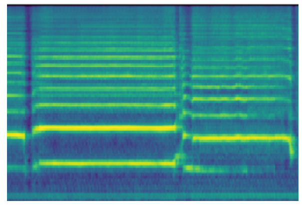
B.2. Attribute Vector vs. Random Noise Conditioning

To address the concern that the observed controllability might stem from arbitrary latent perturbations rather than the proposed $\mathbf{attr}_{\text{inharmonic}}$, we conducted a controlled comparison between the condition using $\mathbf{attr}_{\text{inharmonic}}$ and random noise. We also evaluated the impact of reversed directions via a negative scaling factor m. For a fair comparison, the random noise was generated with the same dimensionality and normalized to have the same L_2 norm as $\mathbf{attr}_{\text{inharmonic}}$.

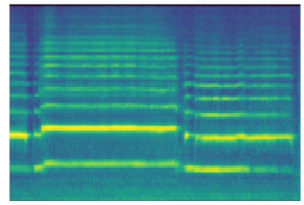
As shown in Table 2, when applying $\mathbf{attr}_{\text{inharmonic}}$ with $m = 5$, both FAD and Pitch Class L_1 increase, while HNR decreases, indicating the emergence of stronger inharmonic partials. Conversely, when applying the inverse attribute vector ($m = -5$), FAD and Pitch Class L_1 decrease, and HNR increases, reflecting a reduction in inharmonicity. In contrast, perturbations with random noise yield metrics nearly identical to the baseline ($m = 0$), regardless of whether $m = 5$ or $m = -5$ is used. The same trend is consistently observed in the Mel spectrogram visualizations shown in Figure 3. Mel spectrograms conditioned on $\mathbf{attr}_{\text{inharmonic}}$ (Figures 3b–3c) exhibit substantially more pronounced differences than those conditioned on random noise (Figures 3d–3e): Figure 3b shows dense inharmonic partials induced by a positive perturbation, whereas Figure 3c becomes noticeably cleaner under the inverse perturbation. These results demonstrate that $\mathbf{attr}_{\text{inharmonic}}$ encodes a meaningful and directional inharmonic transformation, whereas random noise fails to produce comparable effects.



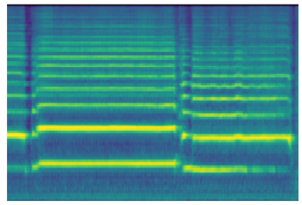
(a) VAE-GAN, $m = 0$



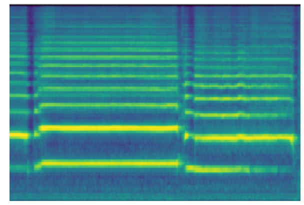
(b) VAE-GAN + Attr, $m = 5$



(c) VAE-GAN + Attr, $m = -5$



(d) VAE-GAN + RN, $m = 5$

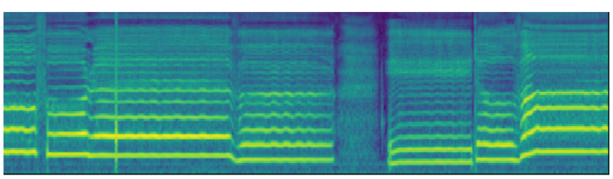


(e) VAE-GAN + RN, $m = -5$

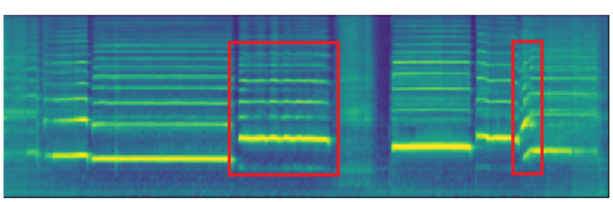
Fig. 3: The Mel spectrograms of the converted output from (a) VAE-GAN with no inharmonic attribute vector or random noise added. (b–c) VAE-GAN with inharmonic attribute vector conditioning at $m = 5$ and $m = -5$. (d–e) VAE-GAN with random noise conditioning at $m = 5$ and $m = -5$. Attr: $\mathtt{attr}_{\text{inharmonic}}$, RN: random noise.

Table 2: Ablation study on inharmonic attribute vector and random noise conditioning. Attr: $\mathtt{attr}_{\text{inharmonic}}$, RN: random noise.

Model	**Cond.**	m	**FAD** ↓	**Pitch Class** L_1	**HNR**
VAE-GAN	–	0	9.09	1.60	25.24
VAE-GAN	Attr	5	11.04	1.62	18.37
VAE-GAN	Attr	-5	**7.64**	**1.59**	28.42
VAE-GAN	RN	5	9.15	1.60	25.10
VAE-GAN	RN	-5	9.11	1.60	25.09



(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 ($\mathcal{L}_{\text{Latent}}$).