

Article

AudioVAE-MASR: A Continuous-Latent Masked Autoregressive Framework for Multi-Distortion Speech Restoration

Fuqiang Hu, Yi Guo *  and Hanbing Tian

School of Electrical Engineering and Electronic Information, Xihua University, Chengdu 610039, China; hfqemail@163.com (F.H.); xhdxthb@163.com (H.T.)

* Correspondence: lpngy@vip.163.com

Featured Application

The proposed framework can be applied to multi-distortion speech restoration, audio inpainting, robust speech communication, and preprocessing for downstream speech systems under degraded acoustic conditions.

Abstract

Real-world speech restoration must handle coupled distortions, including acoustic noise and reverberation, codec artifacts, clipping, and artifacts left by upstream enhancement systems. Token-based generative systems offer a flexible route for such universal restoration, but discrete audio tokens can discard fine acoustic detail, and aggressive generative decoding may over-process inputs that are already close to clean speech. We propose AudioVAE-MASR, a continuous-latent masked autoregressive framework for multi-distortion speech restoration. A frozen AudioVAE maps clean and degraded speech into paired continuous latent sequences; a Conformer-based branch extracts the degraded-condition sequence C_y from degraded latents; a two-stream masked autoregressive encoder-decoder conditions masked clean-latent recovery on both degraded context and visible clean tokens; and a lightweight diffusion head models the masked clean tokens in the continuous latent space. On the released CCF AATC 2025 blind test set, the main inference setting ($K = 16$, temperature 0.5) achieved WAcc 0.793, SIG 3.401, BAK 3.987, OVRL 3.111, PESQ 1.780, and ESTOI 0.798. Relative to the degraded input, these results improved WAcc and DNSMOS but did not improve PESQ; relative to the organizer baseline, they improved WAcc, SIG, OVRL, and PESQ but remained lower in BAK. A local subjective MOS evaluation with five listeners gave an overall mean score of 4.08 for AudioVAE-MASR, compared with 3.70 for the degraded input and 4.59 for the clean reference. Distortion-type, ablation, and parameter-sensitivity analyses further show that codec inputs remain vulnerable to over-restoration and that longer iterative decoding does not provide a consistent gain. The study therefore presents AudioVAE-MASR as a transparent continuous-latent restoration framework and identifies the fidelity-control problems that must be solved before such generative restoration can match the strongest lightweight discriminative systems.

Keywords: speech restoration; continuous latent representation; masked autoregressive; conditional diffusion



Academic Editor: Douglas O'Shaughnessy

Received: 7 May 2026

Revised: 1 July 2026

Accepted: 3 July 2026

Published: 6 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Speech signals are central to telecommunication, human–computer interaction, and downstream speech processing. In practical communication pipelines, however,

speech is rarely corrupted by a single isolated degradation. It may contain non-stationary background noise, room reverberation, clipping, codec artifacts, and secondary artifacts introduced by imperfect upstream enhancement systems [1,2]. Restoring speech under these coupled conditions is difficult because the degradation process is non-linear, the useful speech cues are only partially preserved, and the correct restoration strength depends on the input quality.

Existing speech restoration methods can be broadly categorized into discriminative and generative approaches. Discriminative models optimized with regression-style objectives usually provide stable waveform recovery and strong fidelity, which is reflected by the top systems in the CCF AATC 2025 retrospective [3]. However, regression-oriented models may over-smooth spectra and lose fine perceptual detail [4]. Generative approaches can model richer speech priors, but many systems rely on Vector Quantization (VQ) or aggressive reconstruction in a discrete token space [5,6]. These designs introduce two practical risks for restoration: discretization may discard fine-grained acoustic information, and generative decoding may unnecessarily resynthesize near-clean inputs.

A straightforward cascade of specialized modules, such as denoising followed by dereverberation, is also limited. Such pipelines are vulnerable to error propagation and do not explicitly model the coupling among different distortion sources [7]. This motivates a unified formulation that performs restoration directly under multi-distortion conditions while preserving enough information to diagnose when generative restoration fails.

In this work, we formulate speech restoration as conditional generation in a continuous latent space. The proposed AudioVAE-MASR framework uses a frozen AudioVAE to represent clean and degraded speech as paired continuous latent sequences. A Conformer-based [8] branch extracts the degraded-condition sequence C_y from degraded latents, and a two-stream masked autoregressive encoder–decoder predicts clean-side conditioning features from the degraded stream and visible clean tokens. A lightweight diffusion head then reconstructs masked clean latent tokens in the continuous AudioVAE space. During inference, iterative parallel decoding progressively fills an initially masked clean latent sequence. Using the released CCF AATC 2025 challenge (<https://ccf-aatc.org.cn/> (accessed on 5 July 2025)) blind test set, we evaluate the method under multiple decoding settings and compare the resulting averages with organizer-reported reference values from the official retrospective. The results show that the current model improves WAcc and DNSMOS over the degraded input and exceeds the organizer baseline on several metrics, but still trails the best official systems in PESQ and remains vulnerable to over-restoration on high-quality codec inputs.

Our contributions are summarized as follows:

- Continuous-latent restoration formulation. We formulate multi-distortion speech restoration as conditional generation in a continuous latent space, which avoids the discretization bottleneck of token-based generative restoration and provides a suitable target space for high-fidelity reconstruction.
- Two-stream masked autoregressive architecture. We propose an encoder–decoder design in which a Conformer branch extracts the degraded-condition sequence C_y , visible clean tokens are concatenated with the degraded stream, and decoder-side mask token re-insertion produces clean-side conditions for diffusion-based reconstruction.
- Diffusion-based continuous token reconstruction and transparent evaluation. We couple the masked autoregressive decoder with a lightweight diffusion head for continuous token generation and report blind-test averages, distortion-type diagnostics, ablation results, inference sensitivity, local subjective MOS, and qualitative spectrogram examples to show both the potential and the present limitations of the framework.

2. Related Work

2.1. Three Paradigms of Speech Restoration

Early speech restoration relied heavily on digital signal processing techniques, such as spectral subtraction [9] and Wiener filtering [10], which often struggled with non-stationary noise and non-linear distortions. In the deep learning era, discriminative models such as MP-SENet [11] have become strong practical baselines because they preserve waveform fidelity and intelligibility under supervised objectives. The official CCF AATC 2025 retrospective also shows that lightweight discriminative systems dominate the final ranking, especially in WAcc and PESQ [3]. However, regression-oriented discriminative models may still produce over-smoothed spectra and may not fully recover fine perceptual detail [4]. Generative models address this limitation by using stronger speech priors, but the same priors can introduce reconstruction bias or hallucination when the degraded input is already high quality. This trade-off motivates a continuous-latent generative framework whose behavior must be evaluated not only by selected examples but also by aggregate blind-test diagnostics.

2.2. Latent Representation for Speech

To model high-dimensional audio waveforms effectively, modern approaches typically compress audio into a latent space. A dominant strategy involves Vector Quantization (VQ), exemplified by models like DAC [12] and EnCodec [13], which discretize audio into finite tokens to enable GPT-style autoregression (e.g., AudioLM [14], VALL-E [6]). While effective for synthesis, VQ imposes an information bottleneck: the discretization process inevitably discards fine-grained acoustic details such as phase coherence, subtle prosody and timbral nuances, often resulting in robotic artifacts or quantization noise. Recent visual generation research like DiffMAE [15] and MAR [16] suggests an alternative: modeling data directly in the continuous latent space. By replacing cross-entropy with a Diffusion Loss, these methods avoid quantization errors entirely. Recent speech restoration studies have also begun to explore latent-space generative modeling for high-fidelity reconstruction [17–20]. Our work follows this direction by using a frozen AudioVAE as a continuous analysis/synthesis space. The current experiments do not show that continuous latents alone are sufficient for superior benchmark performance; rather, they allow us to isolate a useful design direction while exposing the need for stronger conditioning and more conservative decoding.

2.3. Masked Autoregressive Modeling

Standard Autoregressive (AR) models (e.g., WaveNet [21], GPT-3 [22]) generate sequences in a strict raster-scan order. While powerful, they are constrained by causality because they cannot utilize future context to refine current predictions and suffer from linear inference latency. Conversely, masked modeling (e.g., BERT [23], MAE [24]) leverages bidirectional context but has traditionally been limited to representation learning rather than generation. Recent hybrid frameworks like MaskGIT [25] and Masked Autoregressive models [16] bridge this gap by employing iterative parallel decoding. In the audio domain, approaches like VampNet [26] have explored masked generation but rely on discrete tokens. Our work applies continuous masked autoregressive modeling to conditional speech restoration. The model differs from strictly causal AR systems by using bidirectional masked context and iterative parallel decoding, and it differs from discrete masked audio generation by predicting continuous AudioVAE latents through a diffusion loss.

3. Materials and Methods

3.1. Overall Framework and Notation

Let $x \in \mathbb{R}^T$ denote a clean speech waveform and let $y \in \mathbb{R}^T$ denote the corresponding degraded waveform. We model the benchmark degradation process as

$$y = \Phi_\delta(x), \quad \delta \in \Delta, \quad (1)$$

where δ denotes the concrete degradation configuration applied to an utterance and Δ denotes the set of possible configurations in the CCF AATC 2025 task, including acoustic degradation, codec distortion, and secondary processing artifacts [3]. The restoration problem is therefore to estimate a clean waveform $\hat{x}$ from y without assuming that the distortion type is known at test time.

Instead of predicting the waveform directly, AudioVAE-MASR performs conditional generation in a continuous latent space. A frozen AudioVAE encoder $\mathcal{E}_A$ maps both waveforms into paired latent sequences:

$$Z_x = \mathcal{E}_A(x), \quad Z_y = \mathcal{E}_A(y), \quad Z_x, Z_y \in \mathbb{R}^{N \times d_z}, \quad (2)$$

where N is the latent sequence length and d_z is the AudioVAE latent dimension. The trainable restoration model learns a conditional latent distribution

$$p_\theta(Z_x | Z_y), \quad (3)$$

where θ denotes all trainable parameters in the Conformer condition extractor, Transformer encoder–decoder, and diffusion head. After sampling the restored latent sequence $\hat{Z}_x \sim p_\theta(Z_x | Z_y)$, the frozen AudioVAE decoder $\mathcal{G}_A$ reconstructs the enhanced waveform:

$$\hat{x} = \mathcal{G}_A(\hat{Z}_x). \quad (4)$$

Thus, the AudioVAE encoder and decoder [27] provide a fixed analysis/synthesis space, while the proposed model learns how to transform degraded continuous latents into clean continuous latents.

The symbols used below follow Figure 1: Z_y and Z_x are the degraded and clean AudioVAE latents, respectively; C_y is the degraded-condition sequence; $\mathcal{M}$ is the set of masked clean positions; and C denotes the clean-side conditioning features passed to the lightweight diffusion head.

3.2. Continuous Latent Representation

AudioVAE [27] is used as a frozen continuous codec rather than as a trainable part of the restoration network. The public implementation is available in the VoxCPM repository (https://github.com/OpenBMB/VoxCPM/blob/main/src/voxcpm/modules/audiovae/audio_vae.py (accessed on 9 October 2025)). In the implementation used in this work, the AudioVAE encoder is a fully convolutional causal encoder with channel expansion and strided downsampling, and the decoder mirrors this structure with transposed convolutional blocks. The configuration uses encoder strides [2, 5, 8, 8], a total hop length of 640 samples, and a 16 kHz sampling rate. Therefore, one latent token corresponds to approximately 40 ms of audio and the latent frame rate is 25 Hz. The latent dimension is $d_z = 64$.

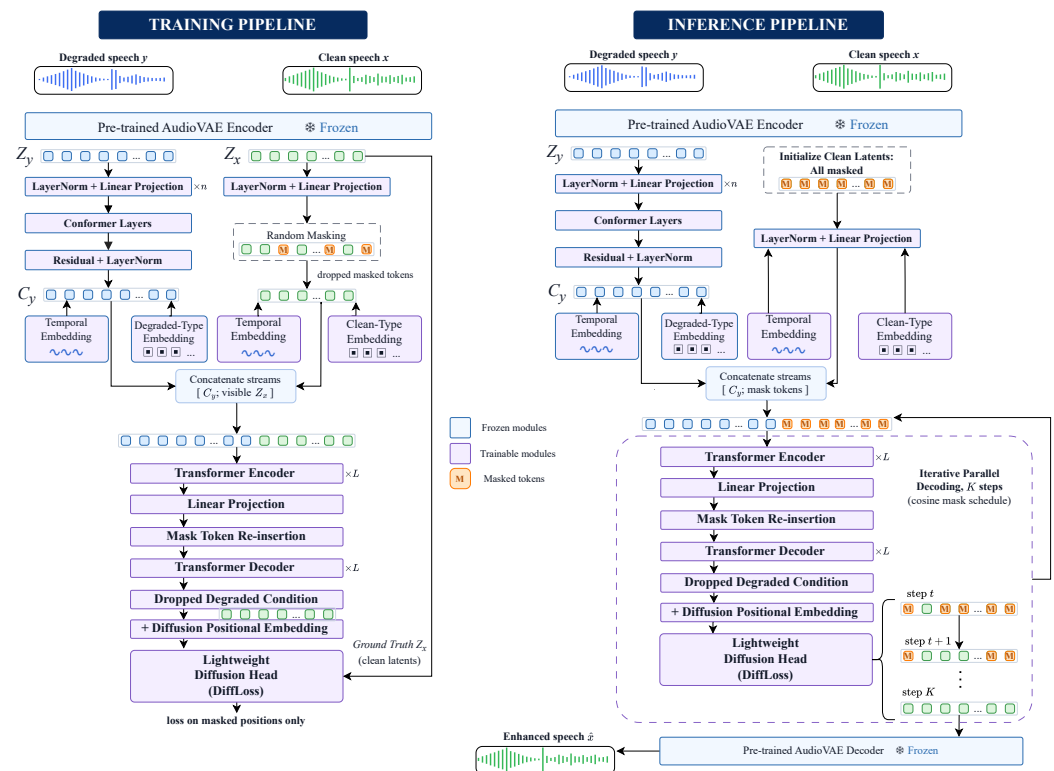


Figure 1. Overview of the proposed AudioVAE-MASR architecture. The arrows indicate the information flow between different modules. During training, clean speech x and degraded speech y are encoded by a frozen AudioVAE encoder into continuous latent sequences Z_x and Z_y . The degraded stream is normalized, linearly projected, and processed by Conformer [8] layers to obtain the degraded-condition sequence C_y . The clean stream is randomly masked; only visible clean tokens are concatenated with C_y and passed through the Transformer encoder. The Transformer decoder restores the full clean-token layout by mask token re-insertion, drops the degraded-condition part, adds diffusion positional embeddings, and sends the resulting clean-side condition to a lightweight diffusion head. The diffusion loss is computed only on masked clean positions. During inference, the clean latent sequence is initialized as fully masked and is progressively filled by iterative parallel decoding under a cosine mask schedule. Frozen and trainable modules are indicated in the figure.

This continuous representation avoids the additional quantization step required by VQ-based audio tokenizers. It also gives the restoration model a compact target space: the model predicts missing or corrupted continuous latent tokens and then relies on the frozen AudioVAE decoder to synthesize the waveform. The decoder used by AudioVAE was pretrained with reconstruction and adversarial objectives, including mel-spectrogram reconstruction and multi-scale/multi-period discriminator losses. In this manuscript, those objectives belong to the pretrained AudioVAE and are not optimized during restoration training.

3.3. Degraded-Condition Extraction

The degraded branch in Figure 1 converts Z_y into a condition sequence C_y before the masked autoregressive backbone. The input degraded latent is first normalized and projected to the Transformer hidden dimension d_h :

$$U_y = W_y \text{LN}(Z_y) + b_y, \quad U_y \in \mathbb{R}^{N \times d_h}. \quad (5)$$

It is then processed by a Conformer stack, followed by a residual connection and LayerNorm:

$$C_y = \text{LN}(\text{Conformer}(\text{LN}(U_y)) + U_y). \quad (6)$$

In our implementation, the Conformer condition extractor contains 6 layers, 8 attention heads, a feed-forward dimension of $4d_h$, and a depthwise convolution kernel size of 31. This design is used because the degraded waveform may contain both long-range effects, such as reverberation and bandwidth limitation, and local transient effects, such as clipping or codec artifacts. The Conformer branch preserves the full degraded sequence; no random masking is applied to Z_y .

Before concatenation with the clean stream, temporal embeddings and degraded-type embeddings are added to C_y . The “degraded-type” and “clean-type” embeddings in Figure 1 are learned stream-type embeddings: they distinguish the degraded-condition stream from the clean-latent stream inside the shared Transformer blocks. They are not oracle labels of the AATC distortion category.

3.4. Masked Two-Stream Autoregressive Backbone

The clean branch starts from Z_x during training. The clean latents are normalized, linearly projected to d_h , and augmented with temporal and clean-type embeddings:

$$H_x = W_x \text{LN}(Z_x) + b_x + P_{\text{enc}} + E_{\text{clean}}, \quad (7)$$

where P_{enc} is the encoder-side temporal embedding and E_{clean} is the clean-type embedding. For each training sample, a random subset of clean positions $\mathcal{M} \subset \{1, \dots, N\}$ is selected. Let $\mathcal{V} = \{1, \dots, N\} \setminus \mathcal{M}$ denote the visible clean positions. The Transformer encoder receives the full degraded-condition stream and only the visible clean tokens:

$$S_{\text{enc}} = [C_y + P_{\text{enc}} + E_{\text{deg}}; H_{x,\mathcal{V}}], \quad (8)$$

where $[\cdot; \cdot]$ denotes sequence concatenation and E_{deg} is the degraded-type embedding. Masked clean tokens are dropped before the encoder, as shown by the “dropped masked tokens” annotation in Figure 1. This keeps the degraded condition always available while forcing the network to infer masked clean latents from the degraded sequence and the remaining visible clean context.

The output of the Transformer encoder is projected to the decoder hidden dimension and re-expanded to the full two-stream layout by mask token re-insertion:

$$S_{\text{dec}} = \text{Reinsert}(W_{\text{dec}} \text{Enc}_{\theta}(S_{\text{enc}}), \mathcal{M}, m) + P_{\text{dec}} + E_{\text{stream}}, \quad (9)$$

where m is the learnable mask token, P_{dec} is the decoder-side temporal embedding, and E_{stream} denotes the corresponding degraded-type or clean-type embedding for each token. The Transformer decoder then updates the complete two-stream sequence:

$$R = \text{Dec}_{\theta}(S_{\text{dec}}). \quad (10)$$

After decoding, the degraded-condition half of R is discarded. The clean-side hidden states are retained and augmented with a learned diffusion positional embedding:

$$C = R_{\text{clean}} + P_{\text{diff}}, \quad C \in \mathbb{R}^{N \times d_h}. \quad (11)$$

These features are not directly projected to clean latents and are not converted to discrete tokens. They serve as token-wise conditioning features for the diffusion head.

3.5. Diffusion Loss with AdaLN Conditioning

The lightweight diffusion head models each clean latent token in the continuous AudioVAE space. For a masked clean position $i \in \mathcal{M}$, let $z_{0,i} = Z_{x,i}$ be the clean target token. The forward diffusion process samples a timestep t and Gaussian noise $\epsilon_i \sim \mathcal{N}(0, I)$:

$$z_{t,i} = \sqrt{\bar{\alpha}_t} z_{0,i} + \sqrt{1 - \bar{\alpha}_t} \epsilon_i. \quad (12)$$

The diffusion network ϵ_θ predicts the noise from the noisy token $z_{t,i}$, timestep t , and decoder condition C_i :

$$\hat{\epsilon}_i = \epsilon_\theta(z_{t,i}, t, C_i). \quad (13)$$

The training loss is applied only to masked positions:

$$\mathcal{L}_{\text{diff}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \mathbb{E}_{t, \epsilon_i} [\|\epsilon_i - \epsilon_\theta(z_{t,i}, t, C_i)\|_2^2]. \quad (14)$$

The implementation uses a cosine noise schedule with 1000 training timesteps and a learned-range variance parameterization; therefore, the final diffusion layer outputs $2d_z$ channels, corresponding to the predicted noise-related channels and variance-related channels used by the diffusion objective.

The conditioning mechanism follows Adaptive Layer Normalization (AdaLN) [28]. The noisy token is first projected to the diffusion hidden width. The timestep embedding and the projected condition are summed:

$$a_i = f_t(t) + W_c C_i. \quad (15)$$

For each residual MLP block, a_i generates shift, scale, and gate vectors:

$$(\beta_i, \gamma_i, g_i) = W_{\text{ada}} \sigma(a_i), \quad (16)$$

and the block update is

$$h_i^{(\ell+1)} = h_i^{(\ell)} + g_i \odot \text{MLP}\left((1 + \gamma_i) \odot \text{LN}(h_i^{(\ell)}) + \beta_i\right). \quad (17)$$

This makes the denoising process explicitly dependent on both diffusion time and the clean-side condition produced by the masked autoregressive backbone. In the configuration used for the reported model, $d_h = 768$, the diffusion hidden width is 1024, and the diffusion MLP depth is 6.

3.6. Random Masking Strategy

AudioVAE-MASR follows the masked generative training strategy of MAR [16], but adapts it to paired degraded/clean speech latents. For each training example, a random generation order π is sampled over the N clean latent positions. The masking ratio is drawn from a left-truncated normal distribution centered near full masking:

$$r \sim \text{TruncNormal}(\mu = 1, \sigma = 0.25; r_{\min} \leq r \leq 1), \quad r_{\min} = 0.7. \quad (18)$$

The number of masked positions is

$$M = \lceil rN \rceil, \quad (19)$$

and the first M positions in π form the mask set $\mathcal{M}$. The diffusion loss above is computed only over $\mathcal{M}$. This training objective encourages the model to use C_y and the visible

clean tokens jointly, rather than learning an unconditional clean-speech prior or a direct point-wise regression from Z_y to Z_x .

3.7. Iterative Parallel Inference

At inference time, only degraded speech y is available. The frozen AudioVAE encoder first produces Z_y , and the degraded-condition branch produces C_y . The clean latent buffer is initialized as a zero-valued sequence but is treated as fully masked, corresponding to the “Initialize Clean Latents: All masked” block in Figure 1. A random order π is sampled once for the sequence, and the model performs K parallel refinement steps.

At step k , the current latent buffer and mask are passed through the same two-stream encoder–decoder path used in training. The decoder produces clean-side conditions $C^{(k)}$ for the currently masked positions, and the diffusion head samples those positions in parallel. The number of positions that remain masked for the next step follows a cosine schedule [16,25]:

$$M_k = \left\lfloor N \cos\left(\frac{\pi k}{2K}\right) \right\rfloor, \quad k = 1, \dots, K. \quad (20)$$

Positions removed from the mask between step $k - 1$ and step k are filled with sampled latent tokens, while the remaining positions continue to use mask tokens at the next decoder pass. The final step fills all remaining masked positions and produces $\hat{Z}_x$. The AudioVAE decoder then maps $\hat{Z}_x$ back to the enhanced waveform $\hat{x}$. In the reported configuration, the default number of iterative decoding steps is $K = 16$, and the diffusion reverse process is respaced to 100 sampling steps for each decoding step.

4. Experimental Setup

4.1. Dataset

We utilized this dataset to rigorously evaluate our model under a challenging multi-distortion speech restoration setting. As illustrated in the challenge guidelines [3], this dataset represents highly complex, real-world scenarios characterized by intertwined impairments. It specifically encompasses three distinct input streams:

1. **Acoustic Degradation:** This stream comprises distorted audio signals degraded using the degradation pipeline introduced by AnyEnhance [29]. The pipeline consists of four stages: reverberation, clipping, bandwidth limitation and additive noise.
2. **Codec Distortion:** This category involves audio signals compressed using MP3 encoding with the torchaudio backend. This process introduces characteristic artifacts such as high-frequency cutoff and quantization noise, particularly at lower bitrates.
3. **Secondary Processing Artifacts:** This set simulates the artifacts introduced by imperfect upstream enhancement [3].

The foundation of the dataset consists of clean speech from three publicly available corpora to ensure diversity in speakers, languages, and recording conditions:

- **VCTK Corpus [30]:** A multi-speaker English dataset with various accents.
- **AISHELL-3 [31]:** A large-scale Mandarin Chinese speech corpus.
- **EARS Dataset [32]:** A high-quality English speech dataset recorded in controlled environments.

Following the official retrospective report [3], the static training set contains 153,475 clean utterances (approximately 300 h). For each clean utterance, the benchmark synthesizes acoustic, codec, and secondary-artifact branches through the three degradation pipelines above. The released development split contains 500 paired utterances sampled from held-out speakers, while the released blind test set contains 300 utterances divided into 150 Acoustic Degradation, 50 Codec Distortion, and 100 Secondary Processing Artifacts examples.

In our experiments, we used the released training manifest, the released development split, and the released blind test set. The training manifest stores the clean path together with the degraded branch and optional auxiliary distorted variants produced by VoiceFixer [33], Demucs [34], FRCRN [35], NSNet2 [36], TF-GridNet [37], Storm [38], SGMSE+ [39], AnyEnhance [29], MaskSR [5], LLaSE-G1 [40], and a codec branch. The released development manifest exposes paired clean/degraded waveforms, whereas the released blind test set is accompanied by metadata containing the official distortion-type labels. Consequently, the distortion-type analysis in this paper is reported on the released blind test set rather than inferred from development-set heuristics.

Throughout the remainder of this paper, “development split” refers to the released development split, whereas “blind test set” refers to the released AATC blind test set. For evaluation, the released AATC blind test set therefore follows the same three official distortion categories used in the organizer report [3]. All audio samples were resampled to 16 kHz to align with the input specifications of AudioVAE.

4.2. Training Setup

The AudioVAE-MASR configuration follows the model code used to generate Figure 1: the Transformer encoder and decoder use hidden dimension $d_h = 768$, 12 attention heads, and 12 blocks each. The degraded-condition branch uses a 6-layer Conformer with 8 attention heads and a depthwise convolution kernel size of 31. The AudioVAE latent dimension is $d_z = 64$. The lightweight diffusion head uses hidden width 1024, depth 6, a cosine noise schedule with 1000 training timesteps [41], and 100 respaced reverse diffusion steps at inference time [16].

The model was optimized with AdamW using an initial learning rate of 5×10^{-5} , $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.02, cosine learning-rate decay, and 2000 warmup steps. Gradient clipping with a maximum norm of 3.0 was applied. The training configuration uses sequence length 50, batch size 72, and a maximum of 100,000 optimization steps. For the blind-test analyses in this manuscript, we evaluate two iterative decoding lengths, $K = 16$ and $K = 32$, and four diffusion sampling temperatures, $\tau \in \{0.5, 0.6, 0.8, 1.0\}$. Unless otherwise stated, the main reported configuration is $K = 16$, $\tau = 0.5$ because it gives the highest overall PESQ and the most conservative quality–intelligibility trade-off among the tested settings.

4.3. Metrics

Following the organizer retrospective [3], we report WAcc, DNSMOS, and PESQ as the main paper-facing metrics. For our generated outputs, WAcc is computed as $1 - \text{WER}$ from the local ASR evaluation output and clipped to $[0, 1]$. Because the competing systems’ restored waveforms are not publicly available, organizer-reported values from the retrospective are used only as contextual reference values; we do not claim an official challenge ranking for our model. We additionally compute ESTOI and P.808 MOS prediction as diagnostic metrics for our own outputs. The P.808 value is a non-intrusive model prediction and should not be interpreted as a subjective listening-test MOS. We employ the following metrics:

1. PESQ (Perceptual Evaluation of Speech Quality) [42]: It is widely adopted for estimating perceptual speech quality relative to a clean reference. PESQ is sensitive to reference-aligned spectral and temporal distortions and is therefore useful for detecting over-restoration.
2. DNSMOS (SIG, BAK, OVRL) [36]: This reference-free estimator is a neural-network-based MOS predictor that correlates with human quality ratings and provides three distinct scores:

- SIG (Signal Quality): Measures the naturalness and quality of the speech component itself. High SIG scores indicate effective restoration of codec artifacts and clipping.
 - BAK (Background Quality): Evaluates the intrusiveness of background noise and the presence of processing artifacts.
 - OVRL (Overall Quality): Represents the global perceptual quality combining both signal and background aspects.
3. P.808 MOS prediction: Provides an additional non-intrusive perceptual quality estimate for our generated outputs.
 4. ESTOI (Extended Short-Time Objective Intelligibility): We use ESTOI as an intrusive intelligibility-oriented diagnostic. It is computed between the clean reference waveform and the restored waveform after resampling to 16 kHz.
 5. WER/WAcc: We transcribe restored speech with a Whisper-large backend [43], compute WER against the reference transcription by edit distance [44], and report WAcc as $1 - \text{WER}$. This local WAcc is used for reproducible analysis of our outputs and may differ from the organizer's exact hidden scoring implementation.
 6. Local subjective MOS: To complement objective metrics, we include a local subjective Mean Opinion Score (MOS) evaluation as a supplementary perceptual assessment. Five listeners rated the clean reference, degraded input, and AudioVAE-MASR output.

5. Experimental Results

5.1. Overall Blind-Test Comparison

Table 1 compares the main AudioVAE-MASR setting with the organizer-reported overall results from the CCF AATC 2025 retrospective [3]. The table is intended as contextual comparison rather than an official leaderboard entry for our system, because the competing systems' restored waveforms and scoring scripts are not available for full re-evaluation. The main configuration of AudioVAE-MASR uses $K = 16$ iterative decoding steps and sampling temperature $\tau = 0.5$.

Table 1. Overall blind-test comparison against organizer-reported CCF AATC 2025 reference values. Official values are taken from the retrospective report [3]. AudioVAE-MASR is evaluated on the released blind test set using the local metric pipeline. Note: $\uparrow$ indicates higher is better.

System	Type	WAcc $\uparrow$	SIG $\uparrow$	BAK $\uparrow$	OVRL $\uparrow$	PESQ $\uparrow$
Clean	Reference	-	3.541	4.126	3.296	-
Noisy	Reference	0.772	2.891	3.019	2.469	1.906
Official baseline	Generative	0.618	3.290	4.050	3.031	1.672
T099	Discriminative	0.811	3.300	4.092	3.054	2.557
T082	Hybrid	0.779	3.466	4.069	3.191	2.165
T002	Discriminative	0.779	3.398	4.070	3.135	2.317
AudioVAE-MASR	Generative	0.793	3.401	3.987	3.111	1.780

The results show a trade-off between perceptual quality estimates and reference-aligned fidelity. Compared with the degraded input, AudioVAE-MASR improves WAcc from 0.772 to 0.793, DNSMOS SIG from 2.891 to 3.401, BAK from 3.019 to 3.987, and OVRL from 2.469 to 3.111. However, PESQ decreases from 1.906 to 1.780, indicating that the generated waveform is not consistently closer to the clean reference despite higher WAcc and DNSMOS. Compared with the official baseline, AudioVAE-MASR improves WAcc, SIG, OVRL, and PESQ, but remains lower in BAK. It also remains below the strongest official systems in PESQ and is not an official leaderboard entry. These observations support the value of continuous-latent masked generation while keeping the performance claim bounded to the released blind-test evaluation.

5.2. Distortion-Type Analysis

Table 2 compares AudioVAE-MASR with the official degraded input, official baseline, and the top five organizer-reported teams for each distortion type. For AudioVAE-MASR, the same main configuration ($K = 16, \tau = 0.5$) is used for all three categories.

Table 2. Distortion-type-wise blind-test comparison. Official reference values are from the CCF AATC 2025 retrospective [3]. Bold values indicate the best value among those official participating teams within each distortion type. Note: $\uparrow$ indicates higher is better.

Distortion Type	System	WAcc $\uparrow$	SIG $\uparrow$	BAK $\uparrow$	OVRL $\uparrow$	PESQ $\uparrow$
Acoustic Degradation	Clean	-	3.526	4.126	3.281	-
	Noisy	0.731	2.408	2.026	1.837	1.306
	Official baseline	0.530	3.221	4.047	2.976	1.505
	T099	0.790	3.218	4.088	2.978	2.337
	T082	0.756	3.491	4.034	3.191	1.960
	T002	0.743	3.375	4.069	3.112	2.108
	T081	0.730	3.313	4.070	3.051	2.165
	T146	0.754	3.116	3.925	2.839	2.331
	AudioVAE-MASR	0.739	3.397	4.036	3.150	1.660
Codec Distortion	Clean	-	3.556	4.110	3.300	-
	Noisy	0.989	3.547	4.091	3.282	4.112
	Official baseline	0.929	3.537	4.099	3.274	2.638
	T099	0.990	3.546	4.109	3.289	4.324
	T082	0.977	3.528	4.127	3.285	3.619
	T002	0.986	3.560	4.109	3.304	4.117
	T081	0.994	3.567	4.129	3.321	4.159
	T146	0.993	3.534	4.091	3.272	4.186
	AudioVAE-MASR	0.962	3.446	3.985	3.127	2.241
Secondary Processing Artifacts	Clean	-	3.554	4.136	3.315	-
	Noisy	0.743	3.254	3.961	2.978	1.758
	Official baseline	0.607	3.312	4.042	3.041	1.489
	T099	0.764	3.330	4.100	3.084	2.110
	T082	0.735	3.368	4.088	3.113	1.808
	T002	0.738	3.320	4.047	3.057	1.788
	T081	0.728	3.353	4.046	3.078	2.013
	T146	0.750	3.298	4.032	3.024	2.115
	AudioVAE-MASR	0.708	3.369	3.933	3.054	1.516

The category-level results clarify where the current system helps and where it remains risky. For Acoustic Degradation, AudioVAE-MASR improves WAcc, SIG, OVRL, and PESQ over both the degraded input and the official baseline, although it remains below the best official systems in WAcc and PESQ and slightly below the baseline in BAK. For Codec Distortion, the degraded signal is already high quality, and AudioVAE-MASR reduces PESQ from 4.112 to 2.241 and WAcc from 0.989 to 0.962, which is consistent with the organizer's observation that generative systems can over-correct high-SNR codec inputs. For Secondary Processing Artifacts, the model improves WAcc, SIG, OVRL, and PESQ relative to the official baseline, but it still reduces WAcc and PESQ relative to the degraded input. These results indicate that the current conditioning and decoding scheme is useful for heavier acoustic corruption but still lacks a reliable preservation mechanism for near-clean or already processed inputs.

5.3. Inference Sensitivity

We next examine whether the current model can be improved by changing the iterative decoding length or diffusion sampling temperature. Table 3 reports all tested settings, and Figure 2 visualizes the same trends for PESQ, WAcc, OVRL, and ESTOI.

Table 3. Inference sensitivity of AudioVAE-MASR on the released AATC blind test set. WAcc is computed as $1 - \text{WER}$. The setting $K = 16, \tau = 0.5$ is used as the main configuration because it gives the highest PESQ and a conservative trade-off across WAcc, DNSMOS, and ESTOI. Note: $\uparrow$ indicates higher is better.

K	τ	WAcc $\uparrow$	SIG $\uparrow$	BAK $\uparrow$	OVRL $\uparrow$	PESQ $\uparrow$	ESTOI $\uparrow$
16	0.5	0.793	3.401	3.987	3.111	1.780	0.798
16	0.6	0.785	3.389	3.976	3.097	1.766	0.789
16	0.8	0.783	3.351	3.971	3.059	1.703	0.788
16	1.0	0.801	3.301	3.903	2.983	1.607	0.787
32	0.5	0.761	3.404	4.016	3.110	1.738	0.781
32	0.6	0.778	3.422	4.012	3.130	1.737	0.793
32	0.8	0.748	3.385	4.003	3.096	1.691	0.781
32	1.0	0.747	3.285	3.845	2.943	1.575	0.781

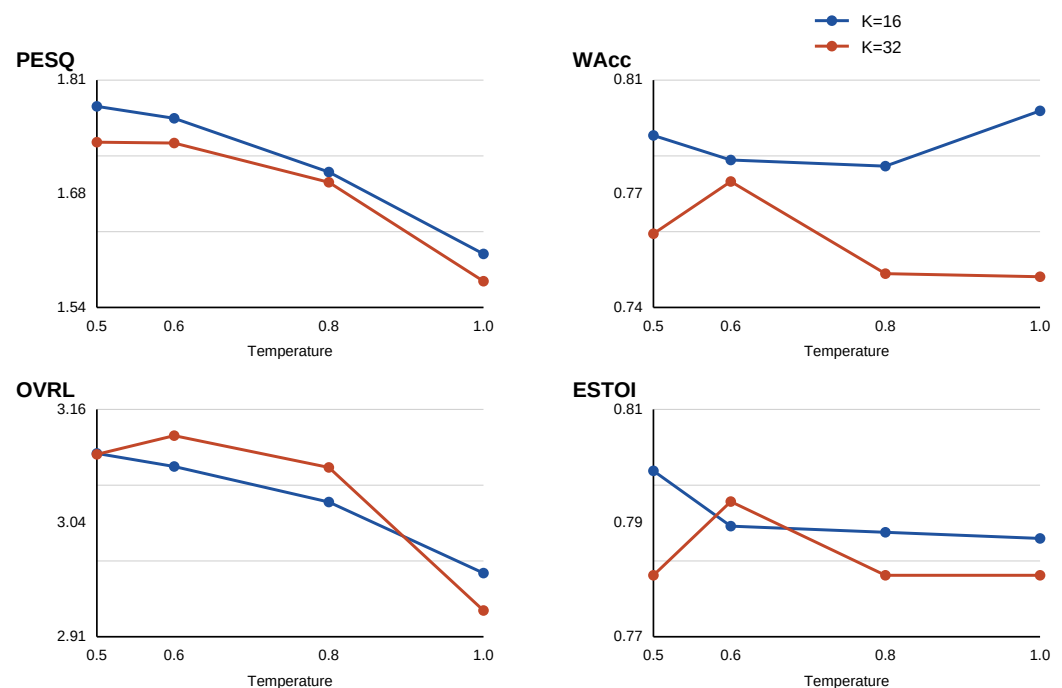


Figure 2. Parameter sensitivity of AudioVAE-MASR on the released AATC blind test set. The curves show that increasing the sampling temperature generally reduces PESQ and DNSMOS, while increasing the number of iterative decoding steps from $K = 16$ to $K = 32$ does not provide a consistent improvement.

The sensitivity analysis shows that the current model is more affected by sampling temperature than by the number of decoding steps. Increasing τ from 0.5 to 1.0 at $K = 16$ decreases PESQ from 1.780 to 1.607 and OVRL from 3.111 to 2.983, even though WAcc changes non-monotonically from 0.793 to 0.801. Changing K from 16 to 32 does not yield a reliable gain: at $\tau = 0.5$, PESQ decreases from 1.780 to 1.738 and WAcc decreases from 0.793 to 0.761, while BAK increases from 3.987 to 4.016. The $K = 32, \tau = 0.6$ setting gives the highest SIG and OVRL, but its PESQ and WAcc remain below the main setting. We therefore treat $K = 16, \tau = 0.5$ as the most conservative setting among those tested.

5.4. Ablation Study

The ablation study isolates the contribution of the main design choices in the final AudioVAE-MASR implementation. Because the architecture centers degraded-condition modeling on the Conformer [8] branch, the third ablation is defined as a Conformer ablation. All ablation variants are evaluated with the same local metric pipeline used for

Tables 1 and 3. The full model is consistently strongest across the ablation groups, but the magnitude of the loss differs by distortion type and by metric.

5.4.1. Effectiveness of Continuous Representations

The first ablation replaces continuous AudioVAE latents with discrete vector-quantized tokens by Encodec [13] while keeping the paired degraded/clean restoration setting unchanged where possible. This comparison tests whether the continuous latent representation reduces quantization-induced information loss for restoration. Compared with the full model, the discrete-token variant lowers WAcc by 0.196 on Acoustic Degradation, 0.378 on Codec Distortion, and 0.050 on Secondary Processing Artifacts. It also reduces OVRL in all three categories. The drop is largest for Codec Distortion, where near-clean inputs require fine acoustic preservation rather than aggressive token-level resynthesis. These results support the use of continuous AudioVAE latents, while also showing that the representation alone does not eliminate over-restoration.

5.4.2. Diffusion Loss vs. MSE Loss

The second ablation replaces the diffusion loss with a direct MSE regression objective on the clean latent targets. This comparison tests whether probabilistic continuous-token reconstruction provides benefits beyond point-wise latent regression. The MSE variant is consistently weaker than the diffusion-loss model in WAcc, OVRL, and ESTOI. The largest degradation appears in Acoustic Degradation, where WAcc decreases from 0.739 to 0.628 and SIG decreases from 3.397 to 3.003. For Codec Distortion, PESQ remains equal to the full model at 2.241, but BAK and ESTOI decrease substantially. This indicates that direct latent regression can preserve some reference-aligned quality in specific cases but does not provide the same balanced restoration behavior across metrics.

5.4.3. Importance of Conformer Condition Extraction

The third ablation removes or replaces the Conformer [8] degraded-condition extractor and therefore tests whether local convolutional modeling plus self-attention in the degraded branch improves restoration under coupled distortions. This replaces the previous decoder-conditioning ablation because the final architecture uses the Conformer degraded-condition branch as the central conditioning mechanism. Removing the Conformer condition extractor produces the largest and most consistent degradation. Relative to the full model, WAcc decreases by 0.360 on Acoustic Degradation, 0.430 on Codec Distortion, and 0.278 on Secondary Processing Artifacts; OVRL decreases by 0.306, 0.755, and 0.382, respectively. The severe drop in Codec Distortion suggests that the condition branch is not only useful for heavy acoustic corruption but also necessary for preserving already informative degraded latents.

5.5. Spectrogram-Based Visual Analysis

Figure 3 provides a visual analysis of the spectral changes introduced by AudioVAE-MASR. The rows were randomly selected from the corresponding blind-test distortion categories rather than manually chosen as favorable cases. The figure is intended to explain typical restoration behavior in randomly sampled examples, while the aggregate conclusions remain based on the quantitative results in Tables 1–4.

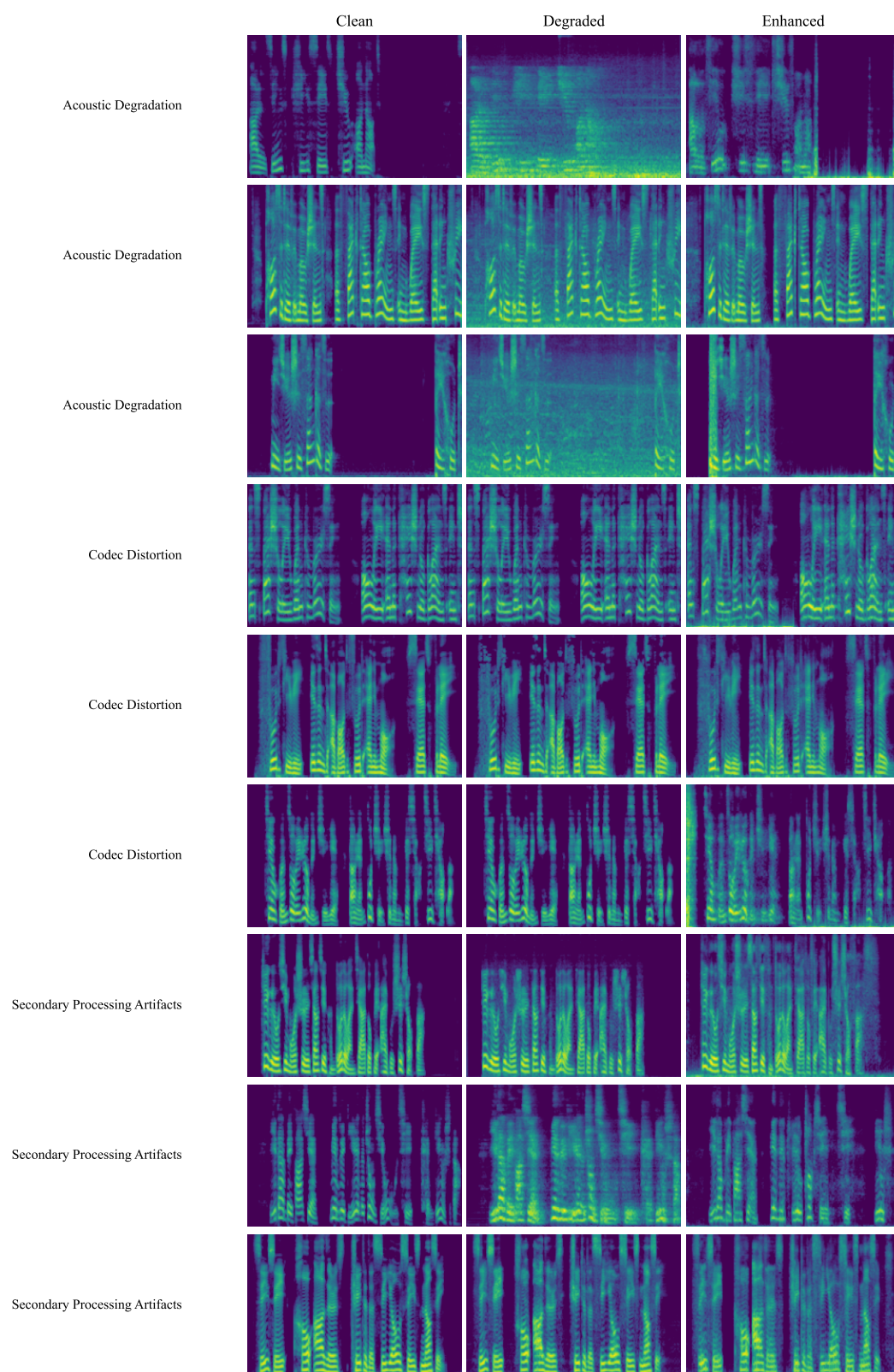


Figure 3. Randomly selected spectrogram examples for AudioVAE-MASR. Each row shows the clean reference, degraded input, and enhanced output for one blind-test utterance, with examples randomly drawn from acoustic degradation, codec distortion, and secondary processing artifacts. The samples are used for qualitative visual analysis of model behavior and should not be interpreted as aggregate superiority over official challenge systems. Color intensity indicates relative spectral energy in the spectrograms.

Table 4. Ablation study for the final AudioVAE-MASR implementation. Full-model values are from the main configuration ($K = 16, \tau = 0.5$), and all variant values are computed with the same local metric pipeline. Note: $\uparrow$ indicates higher is better.

Distortion Type	Variant	WAcc $\uparrow$	SIG $\uparrow$	BAK $\uparrow$	OVRL $\uparrow$	PESQ $\uparrow$	ESTOI $\uparrow$
Acoustic Degradation	AudioVAE-MASR	0.739	3.397	4.036	3.150	1.660	0.773
	Discrete tokens	0.543	3.390	3.846	2.959	1.500	0.755
	MSE loss	0.628	3.003	4.021	3.009	1.564	0.695
	w/o Conformer condition	0.379	2.996	3.549	2.844	1.421	0.595
Codec Distortion	AudioVAE-MASR	0.962	3.446	3.985	3.127	2.241	0.892
	Discrete tokens	0.584	3.412	3.964	2.997	2.117	0.880
	MSE loss	0.873	3.432	3.419	2.957	2.241	0.793
	w/o Conformer condition	0.532	3.051	2.803	2.372	1.936	0.750
Secondary Processing Artifacts	AudioVAE-MASR	0.708	3.369	3.933	3.054	1.516	0.746
	Discrete tokens	0.658	2.966	3.863	2.893	1.498	0.730
	MSE loss	0.700	3.331	3.835	3.001	1.509	0.663
	w/o Conformer condition	0.430	3.147	3.116	2.672	1.380	0.605

For Acoustic Degradation, the degraded spectrograms contain a diffuse broadband noise floor, weakened harmonic ridges, and blurred formant structure. After enhancement, AudioVAE-MASR suppresses much of the low- and mid-frequency background energy and recovers clearer voiced harmonics, making the enhanced spectrograms visually closer to the clean references. This visual trend is consistent with the strong objective and MOS gains observed for Acoustic Degradation, although residual high-frequency texture and occasional over-smoothed regions remain visible.

For Codec Distortion, the degraded inputs are already relatively close to the clean references, so the desirable operation is conservative preservation rather than aggressive regeneration. The enhanced spectrograms show that AudioVAE-MASR can preserve the main harmonic structure, but it also introduces additional smoothing or artificial spectral texture in some regions. This helps explain why Codec Distortion has high intelligibility but reduced PESQ after enhancement: the model changes signals that often require only light correction.

For Secondary Processing Artifacts, the degraded spectrograms show irregular spectral holes, smeared harmonics, and residual artifacts from previous processing. AudioVAE-MASR partially reorganizes harmonic bands and reduces some diffuse artifacts, but several examples still retain unnatural spectral patterns or exhibit newly introduced high-frequency components. The visualization therefore supports the quantitative finding that the method is most reliable under heavy acoustic degradation and less stable when the input has already been codec-processed or enhanced by another system. Because the top-ranked challenge systems do not release their restored waveforms, we do not present side-by-side spectrogram comparisons against T099, T082, or T002.

5.6. Local Subjective MOS Evaluation

Table 5 reports the local subjective MOS results from five listeners. The clean reference, degraded input, and AudioVAE-MASR output were rated by the same five listeners. Because the official final-round MOS values in the CCF AATC 2025 retrospective were obtained by the organizer, the local MOS values in Table 5 should not be directly compared with the official MOS ranking.

Table 5. Local subjective MOS evaluation with five listeners. Values are mean MOS scores for within-study comparison only and are not official CCF AATC 2025 final-round MOS scores.

System	Acoustic Degradation	Codec Distortion	Secondary Processing Artifacts	Overall
Clean reference	4.63	4.54	4.61	4.59
Degraded input	3.04	3.65	4.42	3.70
AudioVAE-MASR	4.11	3.85	4.27	4.08

In the five-listener evaluation, AudioVAE-MASR obtains a higher overall MOS than the degraded input (4.08 vs. 3.70) while remaining below the clean reference (4.59). The subjective gain is strongest for Acoustic Degradation, where MOS increases from 3.04 to 4.11, and modest for Codec Distortion, where MOS increases from 3.65 to 3.85. For Secondary Processing Artifacts, MOS decreases from 4.42 for the degraded input to 4.27 for AudioVAE-MASR, which is consistent with the objective evidence that some already processed inputs can be over-restored. Thus, the local subjective test supports the main conclusion that AudioVAE-MASR improves heavily degraded speech more reliably than near-clean or previously enhanced speech, but it does not establish an official subjective ranking.

5.7. Discussion and Limitations

The revised blind-test results position AudioVAE-MASR as a promising but still bounded continuous-latent generative framework. The model improves WAcc and all three DNSMOS dimensions relative to the degraded input, and it exceeds the official baseline in WAcc, SIG, OVRL, and PESQ. At the same time, it does not improve PESQ over the degraded input and remains below the strongest official systems in PESQ. This behavior is consistent with the organizer retrospective, which reports that lightweight discriminative systems dominate the final ranking and that generative systems can suffer from reconstruction bias under fidelity-oriented metrics [3]. The current results therefore support the representation and modeling motivation, but they do not support an official ranking or a claim of leaderboard-level performance.

The most important failure case is Codec Distortion. In this category, the degraded signal is already close to clean speech according to WAcc, DNSMOS, and PESQ, and restoration should be conservative. AudioVAE-MASR instead reduces PESQ from 4.112 for the degraded input to 2.241 and reduces WAcc from 0.989 to 0.962, indicating over-restoration of high-quality inputs even when intelligibility remains relatively high. The same preservation problem appears, although less cleanly, under Secondary Processing Artifacts, where the input may contain artifacts left by an upstream model rather than simple additive corruption. These results suggest that the current degraded-condition branch and diffusion decoder do not yet estimate when to preserve, lightly edit, or strongly regenerate the signal.

The sensitivity analysis also exposes a practical limitation of the sampling procedure. Higher diffusion sampling temperature increases stochasticity and hurts PESQ and DNSMOS at $\tau = 1.0$, even when WAcc does not follow the same monotonic trend. Increasing the number of iterative parallel decoding steps from $K = 16$ to $K = 32$ does not produce a reliable gain. Thus, the main limitation is unlikely to be solved by simply decoding for more iterations; it requires stronger conditioning, better restoration-strength control, and explicit mechanisms for near-clean input preservation.

The ablation results support the main architectural choices but also clarify their limits. Continuous AudioVAE latents outperform discrete tokens across all distortion categories, and the Conformer condition extractor has the largest module-level effect. However, even the full model over-restores codec inputs, indicating that stronger representation and con-

conditioning are not sufficient without an explicit preservation or bypass mechanism. We also observed that a 1280-dimensional latent representation performed poorly in exploratory trials. This suggests that the current lightweight diffusion head and conditioning pathway do not automatically scale to high-dimensional latent targets, and that high-dimensional continuous representations may require a diffusion architecture specifically adapted to their geometry and noise structure. The local MOS results follow the same pattern: listeners rated AudioVAE-MASR higher than the degraded input overall and under Acoustic Degradation, but the score decreased for Secondary Processing Artifacts. This agreement between objective diagnostics and local subjective evaluation strengthens the conclusion that restoration strength should be input-quality-aware.

Several evaluation limitations remain. First, the official challenge systems are used only as contextual references because their restored waveforms and complete re-evaluation environment are not publicly available. Second, the AudioVAE-MASR values are computed with the local metric pipeline on the released blind test set and should not be read as an official challenge submission. Third, the ablation variants are evaluated locally and should be interpreted as module-level diagnostics rather than independent challenge submissions. Fourth, the local MOS evaluation in Table 5 uses five listeners and is not the official final-round MOS test, so it reduces reliance on objective metrics but does not establish an official subjective ranking. Fifth, broader cross-dataset validation remains necessary. Finally, because the training and evaluation data include speech from multiple language sources, language-specific restoration or adaptation may further improve quality for a target language such as Mandarin Chinese, but this remains outside the scope of the present revision. Future work should therefore add an input-quality-aware bypass or gating module, confidence-estimated restoration strength, stronger discriminative fidelity constraints, diffusion models adapted to high-dimensional latent targets, language-specific adaptation, broader evaluation across datasets, and more standardized subjective or perception-aligned evaluation.

6. Conclusions

This paper presented AudioVAE-MASR, a continuous-latent masked autoregressive framework for multi-distortion speech restoration. The method combines a frozen AudioVAE analysis/synthesis space, a Conformer degraded-condition branch, a two-stream masked autoregressive encoder–decoder, and a lightweight diffusion head for masked clean-latent reconstruction.

Experiments on the released CCF AATC 2025 blind test set show that the main tested setting, using 16 iterative parallel decoding steps ($K = 16$) and a diffusion sampling temperature of 0.5 ($\tau = 0.5$), reaches WAcc 0.793, SIG 3.401, BAK 3.987, OVRL 3.111, PESQ 1.780, and ESTOI 0.798. The model improves WAcc and DNSMOS over the degraded input and exceeds the official baseline in WAcc, SIG, OVRL, and PESQ, but it does not improve PESQ over the degraded input and remains below the strongest official systems in fidelity-oriented comparison. The distortion-type analysis further shows that codec inputs and secondary-artifact inputs remain vulnerable to over-restoration. The ablation study supports the use of continuous latents, diffusion-based reconstruction, and Conformer-based degraded conditioning, while the parameter-sensitivity analysis shows that simply increasing sampling temperature or decoding length does not solve the fidelity-control problem. The local MOS evaluation gives a higher overall score to AudioVAE-MASR than to the degraded input, but it also confirms weaker behavior on already processed inputs. The poor behavior observed with a 1280-dimensional latent representation further suggests that future latent-space restoration models must adapt the diffusion architecture to the target dimensionality rather than only increasing representation size. These findings make the

main contribution of the current study twofold: it specifies a continuous-latent generative restoration architecture and it identifies concrete fidelity-control failure modes that must be solved before such a model can match strong discriminative restoration systems.

Future work will focus on input-quality-aware restoration control, explicit near-clean bypass behavior, stronger fidelity constraints during training, diffusion architectures for high-dimensional latent representations, language-specific adaptation and broader evaluation across datasets.

Author Contributions: F.H. conceived the study, designed the speech restoration framework, implemented the proposed algorithm, conducted the experiments, analyzed the results, and wrote the original manuscript. Y.G. supervised the research, provided guidance on the methodology, experimental design, and manuscript revision, and served as the corresponding author. H.T. contributed to the methodology design, experimental setup, result analysis, and manuscript revision. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed during the current study are publicly available online. The training and development sets of the CCF AATC 2025 dataset can be accessed from the official website at <https://ccf-aatc.org.cn/> (accessed on 5 July 2025). The specific test set utilized for evaluation in this work is openly available in the GitHub repository at https://github.com/viewfinder-annnn/AnyEnhance-v1/releases/tag/v0.1_aatc_testset (accessed on 9 October 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
BAK	Background Quality
CL	Continuous Latent
CNN	Convolutional Neural Network
DDPM	Denoising Diffusion Probabilistic Model
DNSMOS	Deep Noise Suppression Mean Opinion Score
IPD	Iterative Parallel Decoding
MAR	Masked Autoregressive
MOS	Mean Opinion Score
MSE	Mean Squared Error
OVRL	Overall Quality
P.808	ITU-T Recommendation P.808
PESQ	Perceptual Evaluation of Speech Quality
SIG	Signal Quality
SI-SDR	Scale-Invariant Signal-to-Distortion Ratio
SNR	Signal-to-Noise Ratio
STFT	Short-Time Fourier Transform
STOI	Short-Time Objective Intelligibility
VAE	Variational Autoencoder
VQ	Vector Quantization
WAcc	Word Accuracy
WER	Word Error Rate

References

1. Zhang, W.; Scheibler, R.; Saijo, K.; Cornell, S.; Li, C.; Ni, Z.; Kumar, A.; Pirklbauer, J.; Sach, M.; Watanabe, S.; et al. Urgent challenge: Universality, robustness, and generalizability for speech enhancement. *arXiv* **2024**, arXiv:2406.04660.
2. Saijo, K.; Zhang, W.; Cornell, S.; Scheibler, R.; Li, C.; Ni, Z.; Kumar, A.; Sach, M.; Fu, Y.; Wang, W.; et al. Interspeech 2025 URGENT speech enhancement challenge. *arXiv* **2025**, arXiv:2505.23212.
3. Zhang, J.; Zhu, M.; Xu, X.; Bu, H.; Ling, Z.; Wu, Z. The CCF AATC 2025 Speech Restoration Challenge: A Retrospective. *arXiv* **2025**, arXiv:2509.12974.
4. Serbest, S.; Stojkovic, T.; Cernak, M.; Harper, A. DeepFilterGAN: A Full-band Real-time Speech Enhancement System with GAN-based Stochastic Regeneration. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025; pp. 878–882. [\[CrossRef\]](#)
5. Li, X.; Wang, Q.; Liu, X. Maskr: Masked language model for full-band speech restoration. *arXiv* **2024**, arXiv:2406.02092.
6. Chen, S.; Wang, C.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Trans. Audio Speech Lang. Process.* **2025**, *33*, 705–718. [\[CrossRef\]](#)
7. Nair, A.A.; Koishida, K. Cascaded time+ time-frequency unet for speech enhancement: Jointly addressing clipping, codec distortions, and gaps. In *Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2021; pp. 7153–7157.
8. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv* **2020**, arXiv:2005.08100.
9. Berouti, M.; Schwartz, R.; Makhoul, J. Enhancement of speech corrupted by acoustic noise. In *Proceedings of the ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*; IEEE: New York, NY, USA, 1979; Volume 4, pp. 208–211.
10. Lim, J.; Oppenheim, A. All-pole modeling of degraded speech. *IEEE Trans. Acoust. Speech Signal Process.* **1978**, *26*, 197–210. [\[CrossRef\]](#)
11. Lu, Y.X.; Ai, Y.; Ling, Z.H. MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra. *arXiv* **2023**, arXiv:2305.13686.
12. Kumar, R.; Seetharaman, P.; Luebs, A.; Kumar, I.; Kumar, K. High-fidelity audio compression with improved rvqgan. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 27980–27993. [\[CrossRef\]](#)
13. Défossez, A.; Copet, J.; Synnaeve, G.; Adi, Y. High fidelity neural audio compression. *arXiv* **2022**, arXiv:2210.13438.
14. Borsos, Z.; Marinier, R.; Vincent, D.; Kharitonov, E.; Pietquin, O.; Sharifi, M.; Roblek, D.; Teboul, O.; Grangier, D.; Tagliasacchi, M.; et al. Audioldm: A language modeling approach to audio generation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 2523–2533. [\[CrossRef\]](#)
15. Wei, C.; Mangalam, K.; Huang, P.Y.; Li, Y.; Fan, H.; Xu, H.; Wang, H.; Xie, C.; Yuille, A.; Feichtenhofer, C. Diffusion models as masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2023; pp. 16284–16294.
16. Li, T.; Tian, Y.; Li, H.; Deng, M.; He, K. Autoregressive image generation without vector quantization. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 56424–56445. [\[CrossRef\]](#)
17. Zhang, C.; Chen, Z.; Zheng, K.; Zhu, J. VoiceBridge: Designing Latent Bridge Models for General Speech Restoration at Scale. *arXiv* **2025**, arXiv:2509.25275.
18. Jukić, A.; Korostik, R.; Balam, J.; Ginsburg, B. Schrödinger Bridge for Generative Speech Enhancement. In Proceedings of the Interspeech 2024, Kos, Greece, 1–5 September 2024; pp. 1175–1179. [\[CrossRef\]](#)
19. Han, S.; Lee, S.; Lee, J.; Lee, K. Few-step Adversarial Schrödinger Bridge for Generative Speech Enhancement. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025; pp. 2380–2384. [\[CrossRef\]](#)
20. Wang, S.; Liu, S.; Harper, A.; Kendrick, P.; Salzmänn, M.; Cernak, M. Diffusion-based Speech Enhancement with Schrödinger Bridge and Symmetric Noise Schedule. *arXiv* **2024**, arXiv:2409.05116.
21. Van Den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499.
22. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
23. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186.
24. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 16000–16009.
25. Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; Freeman, W.T. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 11315–11325.

26. Garcia, H.F.; Seetharaman, P.; Kumar, R.; Pardo, B. Vampnet: Music generation via masked acoustic token modeling. *arXiv* **2023**, arXiv:2307.04686.
27. Zhou, Y.; Zeng, G.; Liu, X.; Li, X.; Yu, R.; Wang, Z.; Ye, R.; Sun, W.; Gui, J.; Li, K.; et al. VoxCPM: Tokenizer-Free TTS for Context-Aware Speech Generation and True-to-Life Voice Cloning. *arXiv* **2025**, arXiv:2509.24650.
28. Peebles, W.; Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2023; pp. 4195–4205.
29. Zhang, J.; Yang, J.; Fang, Z.; Wang, Y.; Zhang, Z.; Wang, Z.; Fan, F.; Wu, Z. Anyenhance: A unified generative model with prompt-guidance and self-critic for voice enhancement. *arXiv* **2025**, arXiv:2501.15417.
30. Yamagishi, J.; Veaux, C.; MacDonald, K. CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit, Version 0.92; University of Edinburgh; The Centre for Speech Technology Research (CSTR): Edinburgh, UK, 2019. [\[CrossRef\]](#)
31. Shi, Y.; Bu, H.; Xu, X.; Zhang, S.; Li, M. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv* **2020**, arXiv:2010.11567.
32. Richter, J.; Wu, Y.C.; Krenn, S.; Welker, S.; Lay, B.; Watanabe, S.; Richard, A.; Gerkmann, T. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. *arXiv* **2024**, arXiv:2406.06185.
33. Liu, H.; Liu, X.; Kong, Q.; Tian, Q.; Zhao, Y.; Wang, D.; Huang, C.; Wang, Y. Voicefixer: A unified framework for high-fidelity speech restoration. *arXiv* **2022**, arXiv:2204.05841.
34. Defossez, A.; Synnaeve, G.; Adi, Y. Real time speech enhancement in the waveform domain. *arXiv* **2020**, arXiv:2006.12847.
35. Zhao, S.; Ma, B.; Watcharasupat, K.N.; Gan, W.S. FRCRN: Boosting feature representation using frequency recurrence for monaural speech enhancement. In *Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2022; pp. 9281–9285.
36. Reddy, C.K.; Dubey, H.; Gopal, V.; Cutler, R.; Braun, S.; Gamper, H.; Aichner, R.; Srinivasan, S. ICASSP 2021 deep noise suppression challenge. In *Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2021; pp. 6623–6627.
37. Wang, Z.Q.; Cornell, S.; Choi, S.; Lee, Y.; Kim, B.Y.; Watanabe, S. TF-GridNet: Integrating full-and sub-band modeling for speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 3221–3236.
38. Lemerrier, J.M.; Richter, J.; Welker, S.; Gerkmann, T. StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 2724–2737.
39. Richter, J.; Welker, S.; Lemerrier, J.M.; Lay, B.; Gerkmann, T. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 2351–2364. [\[CrossRef\]](#)
40. Kang, B.; Zhu, X.; Zhang, Z.; Ye, Z.; Liu, M.; Wang, Z.; Zhu, Y.; Ma, G.; Chen, J.; Xiao, L.; et al. LLaSE-G1: Incentivizing generalization capability for llama-based speech enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 13292–13305.
41. Nichol, A.Q.; Dhariwal, P. Improved denoising diffusion probabilistic models. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2021; pp. 8162–8171.
42. Rec, I. P.862.2: Wideband Extension to Recommendation p.862 for the Assessment of Wideband Telephone Networks and Speech Codecs; International Telecommunication Union: Geneva, Switzerland, 2005.
43. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2023; pp. 28492–28518.
44. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707–710.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.