SELF-GUIDANCE: TRAINING VQ-VAE DECODERS TO BE ROBUST TO QUANTIZATION ARTIFACTS FOR HIGH-FIDELITY NEURAL SPEECH CODEC

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural speech codecs, predominantly based on Vector-Quantized Variational Autoencoders (VQ-VAEs), serve as fundamental audio tokenizers for speech large language models (SLLMs). However, their reconstruction fidelity is limited by quantization errors introduced during latent space discretization. Existing solutions typically increase model complexity through larger codebooks or hierarchical quantization, which subsequently intensify the modeling challenge for downstream SLLMs. Inspired by the key insight that the codec decoder produces superior output from continuous pre-quantize embeddings, we propose a novel self-guided training mechanism that addresses this problem by enhancing decoder robustness rather than modifying the quantization process. Our method introduces an additional training objective that aligns the decoder's intermediate features when processing both quantized tokens and continuous pre-quantized embeddings through a feature-mapping loss. Extensive experiments on XCodec2 demonstrate that self-guidance consistently improves reconstruction quality across various codebook sizes and quantization techniques (FSQ, SimVQ), achieving state-of-the-art performance for low-bitrate speech codecs. The method requires minimal additional training cost and no inference-time modifications, offering an efficient solution for high-fidelity neural audio coding. Remarkably, our approach enables a 4× reduction in codebook size while maintaining comparable fidelity. Downstream text-to-speech experiments confirm that this reduction significantly improves LLM-based synthesis performance by simplifying the token modeling space.

1 INTRODUCTION

Audio codecs serve as essential tools for audio compression, originally designed to encode continuous audio signals like human speech into sequences of reconstructable discrete codes, enabling efficient data transmission and storage (Wu et al., 2024a). Recently, neural speech codecs, pioneered by SoundStream (Zeghidour et al., 2021) and EnCodec (Défossez et al., 2022), leverage the Vector-Quantized Variational AutoEncoder (VQ-VAE) (Van Den Oord et al., 2017; Esser et al., 2021) architectures to achieve high-fidelity reconstruction at compression ratios significantly exceeding traditional codecs. This breakthrough facilitates the integration of large language models (LLMs) in speech processing and generation, where the discretized audio tokens could be directly adopted in the standard next-token-prediction frameworks of LLMs. Benefiting from large-scale speech modeling with LLMs, numerous studies have advanced downstream tasks, including text-to-speech generation (Wang et al., 2023; Yang et al., 2023b) and interactive multimodal large language models (MLLMs) (Défossez et al., 2024; Zhan et al., 2024).

The transformation from continuous audio to discrete tokens in a VQ-VAE is enabled by a latent vector quantizer. This component maps continuous latent vectors from the encoder to entries in a finite codebook via nearest-neighbor search (i.e., vector quantization) (Van Den Oord et al., 2017; Yu et al., 2021; Mentzer et al., 2023). The corresponding codebook embeddings are then passed to the decoder to reconstruct the audio waveform.

However, quantization is inherently lossy. As noted in prior work (Liu et al., 2024) and confirmed by our preliminary experiments (Section 3.2), the decoder produces higher-fidelity audio when using the

continuous, pre-quantized latents compared to the quantized tokens. This performance gap confirms that quantization error constitutes a major obstacle to high-fidelity reconstruction, as it restricts the information available to the decoder.

To suppress quantization error, existing neural codecs typically employ strategies such as hierarchical quantization with multiple residual codebooks (Zeghidour et al., 2021; Yang et al., 2023a) or simply scaling up the codebook size (Parker et al., 2024; Xin et al., 2024; Wu et al., 2024b; Ye et al., 2025a). While effective for compression, these approaches introduce significant challenges for downstream LLM modeling. Hierarchical codebooks require complex mechanisms to fit within auto-regressive transformer frameworks (Wang et al., 2023; Yang et al., 2023b; Défossez et al., 2024), while a large codebook size exponentially increases the token modeling space, complicating the language modeling task (Ye et al., 2025b).

Thus, reducing quantization error often involves intricate codec designs that inadvertently transfer complexity to downstream LLMs. In this paper, we shift the focus from modifying the quantizer or latent space to enhancing the decoder itself. Our core idea is to **guide the decoder to narrow the output gap between the pre-quantized latent vectors and the quantized tokens**. By aligning the decoder's outputs for these two inputs, we directly mitigate the artifacts introduced by quantization, thereby relieving the quantizer of the sole burden of error elimination.

To this end, we propose a novel learning scheme for VQ-VAE-based codecs, which we call **self-guidance**. During training, the decoder receives both the quantized token embeddings and the continuous pre-quantized latent vectors. We then apply a feature mapping loss between the decoder's intermediate features or outputs for these two paths. This additional objective uses the high-fidelity output from the pre-quantized latents as a target, guiding the decoder to produce similar, high-quality features when driven by the quantized tokens. Consequently, the decoder becomes more robust to quantization artifacts, enhancing the final reconstructed audio's fidelity.

We implement our self-guidance approach on the state-of-the-art single-codebook neural speech codec XCodec2 (Ye et al., 2025b), applying the feature mapping loss to the outputs of the decoder's transformer backbone. Experiments on LibriSpeech show consistent reconstruction improvements across various codebook sizes and quantization techniques (e.g., FSQ, SimVQ). Notably, we achieve comparable reconstruction quality with only a quarter of the original codebook size. The benefits of a smaller codebook are further demonstrated in downstream text-to-speech LLM experiments. Audio samples are available on our demo website.[1]

Our main contributions are as follows:

1. We propose a novel self-guidance mechanism for VQ-VAEs that directs the decoder to mitigate the detrimental effects of quantization error on reconstruction fidelity.
2. We apply self-guidance to the XCodec2 model, achieving state-of-the-art reconstruction performance for low-bitrate speech codecs.
3. Through extensive experiments, we demonstrate that the improvements generalize across different codebook sizes and vector quantization methods.
4. We provide statistical evidence confirming that self-guidance primarily regulates the decoder rather than the encoder.
5. We show that self-guidance reduces the codec's dependency on large codebooks, yielding significant benefits for downstream LLM-based applications.

2 Related works

2.1 Vector quantization

VQ-VAE (Van Den Oord et al., 2017) introduced discrete latent representations for generative models, and VQ-VAE2 (Razavi et al., 2019) enhanced representation richness through hierarchical architectures. VQGAN (Esser et al., 2021) integrated adversarial networks, establishing a fundamental VQ framework for high-quality generative models such as Stable Diffusion (Rombach et al., 2022).

[1] https://sgvqvae.github.io/sgvqvae-demo

Nevertheless, these methods encounter representation collapse when dealing with large codebook sizes, which restricts their scalability.

To tackle this issue, DALL-E (Ramesh et al., 2021) employs Gumbel-Softmax sampling to activate more codes during training, although only a small subset of codes is used for quantization during inference (Zhang et al., 2023). VQGAN-FC (Yu et al., 2021) mitigates collapse by reducing latent dimensionality and applying L2 normalization. Finite scaler quantization (FSQ) (Mentzer et al., 2023) and its variant Look-up free quantization (LFQ) (Yu et al., 2023) project latents to low-dimensional spaces (e.g., binary codes), but this comes at the cost of model capacity, as performance degrades when codebooks are small or collapse is not severe. Recently, VQGAN-LC (Zhu et al., 2024a) and SimVQ (Zhu et al., 2024b) enable stable training with codebook sizes up to 100k by incorporating a linear projector for the codebook.

2.2 Neural codec

In early neural codec model studies, SoundStream (Zeghidour et al., 2021) utilized residual vector quantizers (RVQs) to distribute the codec model's total bitrate across multiple codebooks, preventing codebook size explosion. However, this hierarchical design complicates downstream applications due to the multiple tokens within each frame, necessitating additional flattening or joint modeling.

In recent years, single-codebook codecs have emerged as a simpler and more efficient alternative, demonstrating strong performance at low bitrates (Li et al., 2024; Guo et al., 2024; Ji et al., 2024; Xin et al., 2024; Della Libera et al., 2025). For instance, BigCodec (Xin et al., 2024) employs larger model sizes and advanced learning objectives to achieve high-fidelity audio decoding from a single quantizer of frame rate 80Hz. Despite these advancements, the reconstruction fidelity of BigCodec on perspective metrics significantly degrades at lower frame rates. While high frame rate incurs longer audio token sequences, resulting in a quadratic increase in downstream LLM computation cost, and the language modeling complexity (Wang et al., 2024).

To address this challenge, XCodec (Ye et al., 2025a) and FocalCodec (Della Libera et al., 2025) integrate pretrained self-supervised audio encoders to support the reconstruction performance on perspective metrics. Thanks to the stabilized quantizer like FSQ, TS3Codec (Wu et al., 2024b) and XCodec2 (Ye et al., 2025b) extend the codebook size to over 2^{16} to further boost the model performance, achieving state-of-the-art performance on single-codebook codec of frame rate around 50Hz. However, the drastically extended codebook size poses a significant challenge to the language modeling of downstream LLMs. This issue motivates the development of this paper to explore an approach that relieves the existing codec model's dependency on the large codebook.

3 Preliminary: the effect of quantization in neural codec

3.1 Revisiting the VQ-VAE framework

The Vector-Quantized Variational Autoencoder (VQ-VAE) framework forms the foundation of modern neural audio codecs. As illustrated in Figure 1 (left), the architecture consists of three main components: an encoder, a vector quantizer, and a decoder. The encoder processes an input audio signal $\boldsymbol{x}$ to produce a sequence of continuous latent embeddings $\boldsymbol{z}_e \in \mathbb{R}^{d_e}$, where d_e is the latent dimension. The vector quantizer then maps each embedding in $\boldsymbol{z}_e$ to the nearest entry in a finite codebook $\mathcal{Q} \subset \mathbb{R}^{d_e}$. This operation produces a sequence of quantized token embeddings $\boldsymbol{z}_q$. Finally, the decoder reconstructs the audio signal $\hat{\boldsymbol{x}}$ from $\boldsymbol{z}_q$. During training, gradients are propagated through the non-differentiable quantization operation using straight-through estimation (STE), which copies gradients from $\boldsymbol{z}_q$ directly to $\boldsymbol{z}_e$.

The quantization process inherently introduces error as it projects continuous latent vectors onto a discrete codebook. The quantization error can be quantified as:

$$e_q = \|\boldsymbol{z}_e - \boldsymbol{z}_q\|_2 \tag{1}$$

This error represents the information loss incurred during discretization.

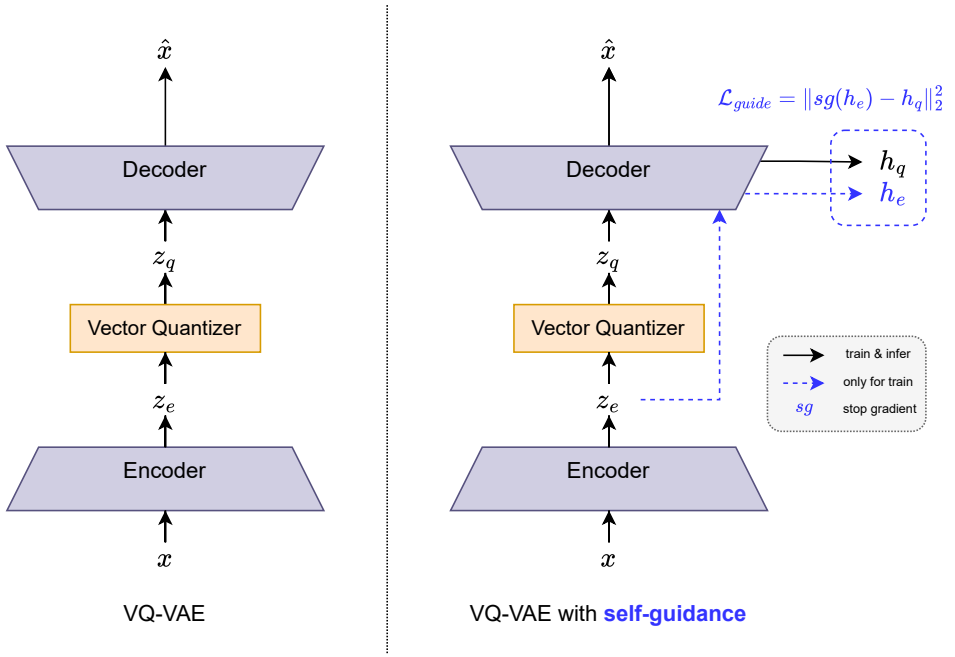

Figure 1: Illustration of the VQ-VAE architecture and the proposed self-guidance (SG) mechanism with the introduced feature mapping loss $\mathcal{L}_{guide}$.

Table 1: Comparing the reconstruction performance of neural speech codec models with different decoder inputs.

Codec model	bitrate	decoder input	STOI↑	WER↓	SIM↑
Ground Truth			1.00	2.4	1.000
Encodec	6kbps	z_q	0.88	2.7	0.861
Encodec	6kbps	z_e	**0.95**	2.7	**0.922**
BigCodec	1.04kbps	z_q	0.93	3.6	0.841
BigCodec	1.04kbps	z_e	**0.95**	**2.9**	**0.872**

3.2 Observation of quantization artifacts

The quantization error e_q introduces information loss that propagates to the decoder, resulting in artifacts that degrade reconstruction fidelity. This phenomenon is evident even though the decoder is exclusively trained on quantized inputs during standard VQ-VAE training.

As shown in Table 1, according to the findings from Liu et al. (2024) on the EnCodec model (Défossez et al., 2022), when the decoder processes the continuous pre-quantized latents $\boldsymbol{z}_e$ instead of the quantized tokens $\boldsymbol{z}_q$, reconstruction quality improves significantly. This observation aligns with our evaluations of BigCodec (Xin et al., 2024).

These results demonstrate that quantization artifacts substantially limit reconstruction quality, presenting a major obstacle for achieving optimal performance in neural codecs.

4 Methodology

To mitigate quantization artifacts in neural speech codecs, we propose a novel learning scheme called **self-guidance** (SG) for VQ-VAE decoders. This section details the self-guidance mechanism and explains our rationale for applying it to the XCodec2 model to construct a high-fidelity neural speech codec.

4.1 SELF-GUIDANCE MECHANISM

The self-guidance mechanism is designed to enhance the decoder's ability to compensate for the information loss caused by quantization error in the input tokens $\boldsymbol{z}_q$. Specifically, we aim to enable the decoder to produce similar outputs from both the quantized tokens $\boldsymbol{z}_q$ and the continuous pre-quantized latents $\boldsymbol{z}_e$.

While the vanilla VQ-VAE reconstruction loss implicitly guides the decoder toward this objective by using the original input $\boldsymbol{x}$ as a target, our preliminary analysis indicates that this alone is insufficient to fully address quantization artifacts. This suggests the need for more explicit guidance during training.

Inspired by our preliminary findings, we propose using the pre-quantized latent $\boldsymbol{z}_e$ itself as an internal guidance signal. As illustrated in Figure 1 (right), during training we introduce an additional forward pass that feeds $\boldsymbol{z}_e$ to the decoder. We then extract intermediate hidden features from both paths: $\boldsymbol{h}_e$ from the $\boldsymbol{z}_e$ branch and $\boldsymbol{h}_q$ from the $\boldsymbol{z}_q$ branch. We introduce a feature-mapping loss $\mathcal{L}_{\text{guide}}$ to align these features:

$$\mathcal{L}_{\text{guide}} = \|\text{sg}(\boldsymbol{h}_e) - \boldsymbol{h}_q\|_2^2 \tag{2}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. This loss term is added to the original VQ-VAE objectives to form an end-to-end self-supervised training process.

The self-guidance mechanism introduces minimal computational and architectural overhead:

- **Training:** Only an additional forward pass through the decoder with $\boldsymbol{z}_e$ is required, with no gradient computation needed for this branch.
- **Inference:** No modifications are required; the decoder operates exclusively on $\boldsymbol{z}_q$ as in standard VQ-VAE models.

4.2 NEURAL SPEECH CODEC MODEL

To validate the effectiveness of self-guidance, we apply it to XCodec2, a state-of-the-art neural speech codec that has demonstrated strong performance in low-bitrate speech encoding and downstream speech generation tasks (Boson AI, 2025; Ye et al., 2025b).

XCodec2 comprises several key components: a convolutional encoder, a single-layer finite scalar quantizer (FSQ), and an acoustic decoder. Additionally, it includes a semantic encoder and decoder that form an auxiliary autoencoder operating on Wav2Vec2-BERT features (Barrault et al., 2023), enhancing the semantic content of the encoded latents for improved downstream performance.

A distinctive feature of XCodec2 is its acoustic decoder architecture. Like in TS3Codec (Wu et al., 2024b), rather than using stacked convolutional upsampling blocks, it employs a Transformer backbone followed by an inverse short-time Fourier transform (iSTFT) head (Siuzdak, 2024). This design naturally suggests using the Transformer backbone outputs for computing $\mathcal{L}_{\text{guide}}$ because: (i) the Transformer contains the majority of learnable parameters in the decoder, providing sufficient capacity to benefit from self-guidance; and (ii) the subsequent iSTFT head separates the hidden features from the final waveform generation, preventing potential interference from waveform-level reconstruction losses.

The complete training objective for our enhanced codec is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{guide}} + \mathcal{L}_{\text{semantic}} + \mathcal{L}_{\text{acoustic}} + \mathcal{L}_{\text{adv}} \tag{3}$$

where:

- $\mathcal{L}_{\text{guide}}$ is the self-guidance feature mapping loss defined in Equation 2, computed using the Transformer backbone outputs;
- $\mathcal{L}_{\text{semantic}}$ is the mean squared error semantic feature reconstruction loss;
- $\mathcal{L}_{\text{acoustic}}$ is the multi-scale mel-spectrogram reconstruction loss;
- $\mathcal{L}_{\text{adv}}$ is the adversarial loss from a multi-period discriminator (Kong et al., 2020) and a spectrogram discriminator (Parker et al., 2024).

5 EXPERIMENTS AND ANALYSIS

5.1 EXPERIMENT SETTINGS

Dataset We use the full Librispeech Panayotov et al. (2015) training set for the training of all versions of codec models, which comprises 960 hours of English speech audio at a sampling rate of 16kHz. For evaluation, the *test-clean* subset of LibriSpeech that contains 2620 utterances from 40 speakers is used to assess reconstruction performance.

Implementation details We build our neural codec model based on the offical open-source code of XCodec2 [2]. The modifications required to implement self-guidance are minimal: (i) adding an additional forward pass in the `forward` function [3]; (ii) incorporating the computation of $\mathcal{L}_{guide}$ in the `compute_gen_loss` function [4]. The full modified code script is attached in the supplementary material. Detailed configurations are included in Section A.2. The BigCodec model involved in the preliminary study (Section 3.2) and comparative experiments (Section 5.2) is obtained via training with the official open-source implementation [5]

Training cost We train all of the codec models on 8 NVIDIA GeForce RTX 4090 GPUs for 600 thousand iterations. The total training time of each codec model is around 237.75 hours. Notably, the self-guidance variant incurs negligible additional training time compared to the baseline XCodec2, with differences of only seconds. This efficiency aligns with our design in Section 4.1: the additional forward pass through the acoustic decoder requires no backward propagation, making the computational overhead minimal compared to other components (e.g., discriminators) and gradient synchronization. This demonstrates that the performance gains from self-guidance come at virtually no additional training cost.

Table 2: Comparing reconstruction evaluation results with other existing neural codecs on the LibriSpeech test-clean dataset. (**SG** signifies the proposed self-guidance mechanism; details about each metric are included in Section A.1)

Codecs models	Frame rate	Codebook size(s)	PESQ↑	STOI↑	MCD↓	WER↓	SIM↑	UTMOS↑
Ground Truth			4.64	1.000	0.00	2.5	1.00	4.08
DAC	50Hz	1024×8	2.72	0.940	–	–	0.87	–
DAC	50Hz	1024×2	1.13	0.730	–	–	0.32	–
WavTokenizer	75Hz	4096	2.05	0.886	4.00	6.8	0.59	3.89
BigCodc	80Hz	8192	2.68	0.935	2.93	3.6	0.84	4.11
WavTokenizer	40Hz	4096	1.88	0.868	4.32	8.0	0.57	3.77
BigCodec	40Hz	8192	<u>2.11</u>	<u>0.894</u>	<u>3.72</u>	6.7	0.66	4.05
XCodec2	50Hz	8192	2.03	0.892	3.84	<u>4.1</u>	<u>0.72</u>	**4.09**
XCodec2+**SG**	50Hz	8192	**2.13**	**0.898**	**3.60**	**3.8**	**0.73**	<u>4.08</u>
TS3Codec	40Hz	65536	2.01	0.893	3.81	4.9	0.61	3.69
TS3Codec	40Hz	131072	2.06	0.897	3.75	4.5	0.63	3.73
TS3Codec	50Hz	65536	2.22	0.909	3.52	3.6	0.68	3.85
TS3Codec	50Hz	131072	2.23	0.910	3.50	3.6	0.68	3.84
XCodec2	50Hz	65536	2.28	0.910	3.57	3.2	0.79	4.06
XCodec2+**SG**	50Hz	65536	**2.39**	**0.915**	**3.41**	**3.2**	**0.80**	**4.10**

[2] https://github.com/zhenye234/X-Codec-2.0

[3] https://github.com/zhenye234/X-Codec-2.0/blob/main/lightning_module.py#L146

[4] https://github.com/zhenye234/X-Codec-2.0/blob/main/lightning_module.py#L239

[5] https://github.com/Aria-K-Alethia/BigCodec

5.2 RECONSTRUCTION PERFORMANCE

We first evaluate the overall reconstruction performance of our proposed model against existing low-bitrate speech codecs. For DAC, WavTokenizer, and TS3Codec, we report results from papers. For BigCodec and XCodec2, we retrain models and include variations with different configurations (XCodec2 default: frame rate = 50 Hz, $|\mathcal{Q}|$ = 65,536; BigCodec default: frame rate = 80 Hz, $|\mathcal{Q}|$ = 8,192).

As shown in Table 2, our proposed model (XCodec2 with self-guidance) achieves the best performance across most evaluation metrics. For codecs with frame rates of 40–50 Hz, our approach consistently outperforms competitors with similar codebook sizes (8,192 and below, or 65,536 and above), establishing new state-of-the-art performance for low-bitrate speech codecs.

Specifically, while the original XCodec2 with codebook size 65,536 shows competitive performance, self-guidance provides further improvements across all metrics. Reducing XCodec2's codebook size to 8,192 significantly degrades acoustic reconstruction quality (PESQ, STOI, MCD), falling behind BigCodec. However, when augmented with self-guidance, this reduced-size model surpasses BigCodec on all metrics

5.3 ABLATION STUDIES

We conduct ablation studies to isolate the contribution of self-guidance and evaluate its robustness under different configurations. We compare models trained with and without self-guidance while varying quantizer settings.

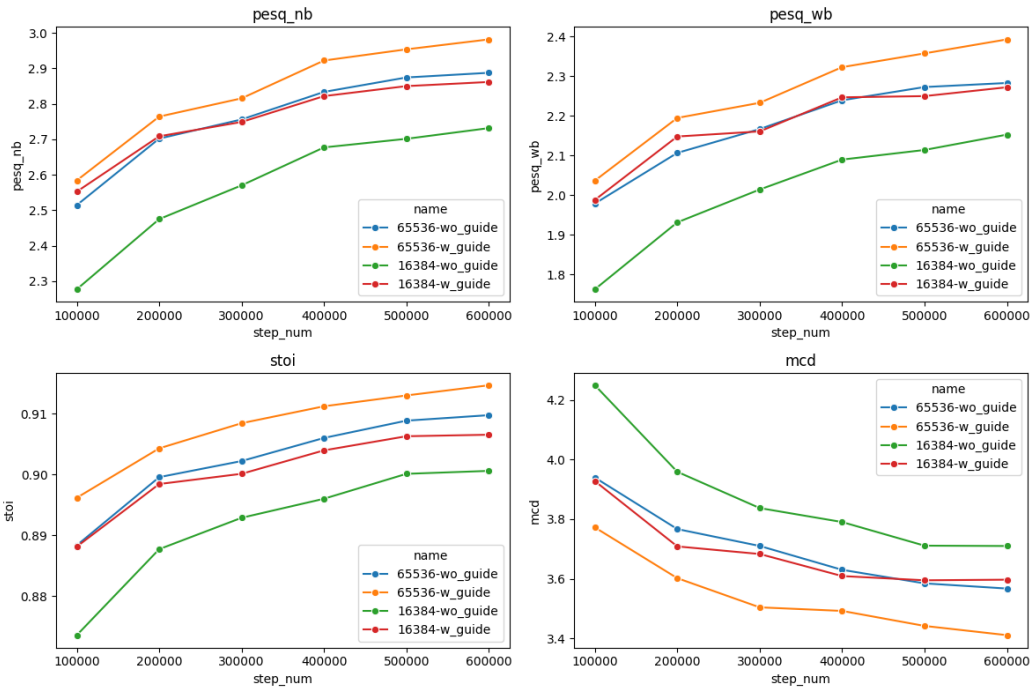

Figure 2: Comparison of the reconstruction performance under various settings along the training process. Horizontal axis is the training iterations. Best viewed in color.

Codebook size We experiment with codebook sizes of 8,192, 16,384, and 65,536. As shown in Table 3, self-guidance improves performance across all settings, except for a minor degradation in word error rate (WER) at the intermediate size (16,384). Figure 2 shows that with self-guidance, a model with codebook size 16,384 achieves similar performance to the baseline XCodec2 with a 4× larger codebook (65,536) on several metrics.

Table 3: Reconstruction evaluation results of the proposed neural speech codec across different codebook sizes. (**with SG** signifies whether the proposed self-guidance mechanism is applied)

Codebook size	with SG	PESQ-WB↑	PESQ-NB↑	STOI↑	MCD↓	WER↓	SIM↑	UTMOS↑
Ground Truth		4.64	4.54	1.000	0.00	2.49	1.00	4.08
8192	✗	2.03	2.59	0.892	3.84	4.08	0.72	**4.09**
8192	✓	**2.13**	**2.69**	**0.898**	**3.79**	**3.77**	**0.73**	4.08
16384	✗	2.15	2.73	0.901	3.71	**3.47**	0.76	3.98
16384	✓	**2.27**	**2.86**	**0.907**	**3.70**	3.53	**0.77**	**4.08**
65536	✗	2.28	2.89	0.910	3.57	3.23	0.79	4.06
65536	✓	**2.39**	**2.98**	**0.915**	**3.41**	**3.15**	**0.80**	**4.10**

Table 4: Reconstruction evaluation results of the proposed neural speech codec across different types of vector quantizers (XCodec2 adopts FSQ by default), with codebook size fixed at 16384. (**with SG** signifies whether proposed self-guidance mechanism is applied).

Quantizer	with SG	PESQ-WB↑	PESQ-NB↑	STOI↑	MCD↓	WER↓	SIM↑	UTMOS↑
Ground Truth		4.64	4.54	1.000	0.00	2.49	1.00	4.08
FSQ	✗	2.15	2.73	0.901	3.71	**3.47**	0.76	3.98
FSQ	✓	**2.27**	**2.86**	**0.907**	**3.60**	3.53	**0.77**	**4.08**
SimVQ	✗	2.10	2.67	0.900	3.63	**3.59**	0.75	3.85
SimVQ	✓	**2.17**	**2.74**	**0.904**	**3.56**	3.63	**0.76**	**3.93**

Type of vector quantizer To assess generalization across quantizer types, we replace the default FSQ quantizer in XCodec2 with SimVQ (VQGAN-FC suffered from codebook collapse and produced unintelligible results). Table 4 shows that self-guidance consistently improves performance with SimVQ, reproducing the minor WER degradation observed with FSQ. We hypothesize this effect relates to self-guidance being applied only to the acoustic decoder.

Quantization error Since $\mathcal{L}_{\text{guide}}$ gradients propagate to the encoder via straight-through estimation, we investigate whether performance gains stem from decoder guidance or implicit quantization error reduction. Figure 3 shows quantization error (e_q) distributions on the test-clean dataset for baseline and self-guidance models across different codebook sizes. The closely overlapping distributions demonstrate that **self-guidance enhances reconstruction by improving decoder robustness** rather than reducing quantization error.

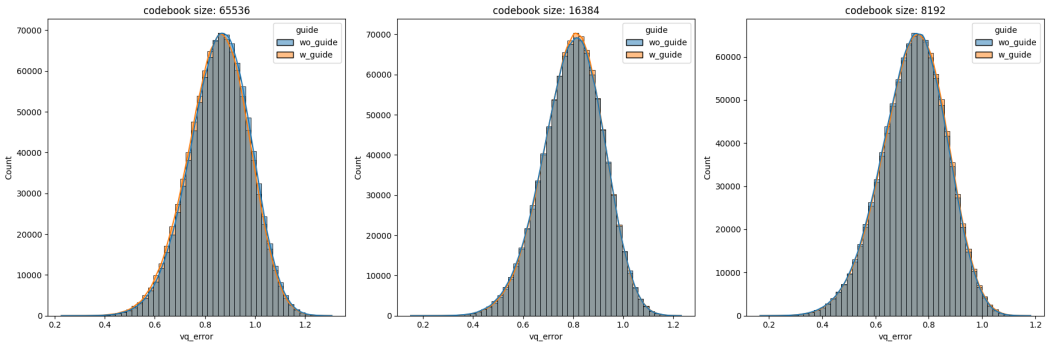

Figure 3: The histogram of the quantization error e_q on LibriSpeech test-clean dataset with the self-guidance mechanism activated (`w_guide`) or omitted (`wo_guide`) across different codebook sizes (from left to right: 65536, 16384, 8192).

5.4 DOWNSTREAM AUTO-REGRESSIVE TTS

Building on our finding that self-guidance enables smaller codebooks to achieve performance comparable to larger ones (Figure 2), we evaluate its impact on downstream autoregressive text-to-speech (TTS) synthesis. We hypothesize that reduced codebook size simplifies the language modeling task, potentially improving final TTS quality.

We conduct rapid TTS experiments using a Qwen2.5-0.5B causal LLM backbone (Qwen et al., 2025) trained on LibriTTS-R (Koizumi et al., 2023). Input text is phonemized before training. Models are supervised fine-tuned for phoneme-to-audio-token generation for 85 epochs. For inference, we use the continual synthesis approach from VALL-E (Wang et al., 2023), providing phoneme sequences and first 3-second audio tokens as prompts for continuation generation. We filter LibriTTS-R test-clean samples longer than 6 seconds and generate continuations using top-k (50) and top-p (0.9) sampling.

As shown in Table 5, results support our hypothesis: the self-guidance model with smaller codebook demonstrates stronger performance in autoregressive audio generation. This indicates that self-guidance not only enhances codec reconstruction fidelity but also facilitates downstream LLM applications by reducing language modeling complexity through smaller codebook requirements.

Table 5: Downstream text-to-speech continuation performance on the LibriTTS test-clean dataset.

Codec model	Codebook size	UTMOS↑	WER↓	SIM↑
XCodec2	65536	3.33	33.03	0.58
XCodec2+**SG**	16384	**3.58**	**28.02**	0.58

6 CONCLUSION

We proposed self-guidance, a novel training mechanism for VQ-VAE-based neural speech codecs that enhances decoder robustness to quantization artifacts. By aligning the decoder's outputs for quantized and continuous latent representations through an additional feature-mapping loss, our method improves reconstruction fidelity without modifying the inference process. Experiments demonstrate that self-guidance consistently enhances performance across various codebook sizes and quantization techniques, enabling comparable quality with 4× smaller codebooks. Downstream TTS results confirm that this reduction simplifies language modeling for LLMs, improving synthesis quality. Our approach provides an effective and efficient solution to mitigate quantization errors, advancing high-fidelity neural audio compression.

Future work could explore applying self-guidance to other VQ-VAE domains beyond speech, such as music or general audio processing, and investigating its combination with more advanced quantization techniques.

Ethics statement This work presents research on neural audio codecs, which have significant potential for positive applications in speech compression, communication, and generative modeling. However, we acknowledge several ethical considerations:

1. **Positive Impacts:** Our method enables higher-quality audio compression at lower bitrates, which could improve accessibility and efficiency in telecommunication, hearing assistance devices, and low-bandwidth applications. The reduction in codebook complexity may also decrease computational requirements for downstream applications.
2. **Potential Misuse:** Like other audio generation technologies, neural codecs could potentially be misused for creating deepfake audio or other deceptive content. However, our work focuses specifically on reconstruction quality rather than generative capabilities. The codec itself does not generate novel content without being integrated into a full generative system.

Reproducibility statement As stated in Section 5.1, the proposed approach is adapted from the official open-source code of XCodec2. The modified code script that implements self-guidance is attached in the supplementary material. We have attached the modified code script to the supplementary materials. We have also described the computational requirements of our experiments in Section 5.1. Details about the model configuration are included in Section A.2 for confirmation.

REFERENCES

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023.

Boson AI. Higgs Audio V2: Redefining Expressiveness in Audio Generation. `https://github.com/boson-ai/higgs-audio`, 2025. GitHub repository. Release blog available at `https://www.boson.ai/blog/higgs-audio-v2`.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518, 2022.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Luca Della Libera, Francesco Paissan, Cem Subakan, and Mirco Ravanelli. Focalcodec: Low-bitrate speech coding via focal modulation networks. *arXiv preprint arXiv:2502.04465*, 2025.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Yiwei Guo, Zhihan Li, Chenpeng Du, Hankun Wang, Xie Chen, and Kai Yu. Lscodec: Low-bitrate and speaker-decoupled discrete speech codec. *arXiv preprint arXiv:2410.15764*, 2024.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*, 2024.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. Libritts-r: A restored multi-speaker text-to-speech corpus. *arXiv preprint arXiv:2305.18802*, 2023.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

Hanzhao Li, Liumeng Xue, Haohan Guo, Xinfa Zhu, Yuanjun Lv, Lei Xie, Yunlin Chen, Hao Yin, and Zhifei Li. Single-codec: Single-codebook speech codec towards high-performance speech generation. *arXiv preprint arXiv:2406.07422*, 2024.

Alexander H. Liu, Qirui Wang, Yuan Gong, and James R. Glass. Closer look at neural codec resynthesis: Bridging the gap between codec and waveform generation. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024. URL `https://openreview.net/forum?id=eO4DmksTv8`.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Julian D Parker, Anton Smirnov, Jordi Pons, CJ Carr, Zack Zukowski, Zach Evans, and Xubo Liu. Scaling transformers for low-bitrate high-quality speech coding. *arXiv preprint arXiv:2411.19842*, 2024.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pp. 749–752. IEEE, 2001.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vY9nzQmQBw.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on audio, speech, and language processing*, 19(7):2125–2136, 2011.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

Hankun Wang, Haoran Wang, Yiwei Guo, Zhihan Li, Chenpeng Du, Xie Chen, and Kai Yu. Why do speech language models fail to generate semantically coherent outputs? a modality evolving perspective, 2024. URL https://arxiv.org/abs/2412.17048.

Haibin Wu, Ho-Lam Chung, Yi-Cheng Lin, Yuan-Kuei Wu, Xuanjun Chen, Yu-Chi Pai, Hsiu-Hsuan Wang, Kai-Wei Chang, Alexander H Liu, and Hung-yi Lee. Codec-superb: An in-depth analysis of sound codec models. *arXiv preprint arXiv:2402.13071*, 2024a.

Haibin Wu, Naoyuki Kanda, Sefik Emre Eskimez, and Jinyu Li. Ts3-codec: Transformer-based simple streaming single codec. *arXiv preprint arXiv:2411.18803*, 2024b.

Detai Xin, Xu Tan, Shinnosuke Takamichi, and Hiroshi Saruwatari. Bigcodec: Pushing the limits of low-bitrate neural speech codec. *arXiv preprint arXiv:2409.05377*, 2024.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *CoRR*, abs/2305.02765, 2023a. URL https://doi.org/10.48550/arXiv.2305.02765.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*, 2023b.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, et al. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25697–25705, 2025a.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, et al. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*, 2025b.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

Jiahui Zhang, Fangneng Zhan, Christian Theobalt, and Shijian Lu. Regularized vector quantization for tokenized image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18467–18476, 2023.

Lei Zhu, Fangyun Wei, Yanye Lu, and Dong Chen. Scaling the codebook size of vqgan to 100,000 with a utilization rate of 99%. *arXiv preprint arXiv:2406.11837*, 2024a.

Yongxin Zhu, Bocheng Li, Yifei Xin, and Linli Xu. Addressing representation collapse in vector quantized models with one linear layer. *arXiv preprint arXiv:2411.02038*, 2024b.

A Appendix

A.1 Evaluation metrics of reconstruction

We evaluate acoustic fidelity, intelligibility, and naturalness of the speech audio reconstructed by neural codecs using the following metrics:

Perceptual Evaluation of Speech Quality (PESQ) PESQ Rix et al. (2001) compares degraded and reference speech to predict human-perceived quality. We use a Python implementation [6] to compute wide-band (PESQ-WB) and narrow-band (PESQ-NB) scores, where higher values indicate better quality.

Mel Cepstral Distortion (MCD) MCD measures the difference between mel-frequency cepstral coefficients (MFCCs), a standard metric for speech synthesis quality.

Short-Time Objective Intelligibility (STOI) STOI Taal et al. (2011) evaluates speech intelligibility by comparing temporal envelopes of clean and degraded signals, with scores ranging from 0 (unintelligible) to 1 (perfect intelligibility).

[6] https://github.com/ludlows/PESQ

Word Error Rate (WER) WER is calculated using a HuBERT Hsu et al. (2021) speech recognition model finetuned on Librispeech [7], reporting percentage errors in transcribed words.

Speaker Similarity (SIM) Speaker characteristics are evaluated via cosine similarity between original and reconstructed utterances, using a WavLM-large Chen et al. (2022)-based speaker verification model [8].

UTMOS UTMOS Saeki et al. (2022) predicts Mean Opinion Score (MOS) for speech naturalness, with scores from 1 (poor) to 5 (excellent). We use a pretrained UTMOS strong model [9].

Table 6: Model configurations

Configuration entry	Value
Acoustic encoder hidden dim	1024
Acoustic encoder convoution blocks	5
Acoustic encoder up ratio	[2, 2, 4, 4, 5]
Acoustic decoder hidden dim	1024
Acoustic decoder Transformer layers	12
Semantic encoder hidden dim	1024
Semantic decoder hidden dim	1024
FSQ scales (codebook size = 65536)	[4, 4, 4, 4, 4, 4, 4, 4]
FSQ scales (codebook size = 16384)	[4, 4, 4, 4, 4, 4, 4]
FSQ scales (codebook size = 8192)	[4, 4, 4, 4, 4, 4, 2]
loss weight $\lambda_{semantic}$	5.0
loss weight $\lambda_{acoustic}$	15.0
loss weight λ_{adv}	1.0
loss weight λ_{guide} (codebook size = 65536)	5.0
loss weight λ_{guide} (codebook size = 16384)	10.0
loss weight λ_{guide} (codebook size = 8192)	10.0
batch size	16
optimizer	AdamW
optimizer betas	[0.8, 0.9]
learning rate warmup steps	1000
learning rate decay steps	500000
learning rate min value	2e-5
learning rate max value	1e-4

A.2 Model configuration

The detailed model configuration and loss weights are listed in Table 6. Most of the configurations follows the default configuration of XCodec2. Specifically, the weight of the proposed self-guidance feature mapping loss weight λ_{guide} is selected from the best of [1, 5, 10, 15], according to the overall reconstruction performance in test trials.

B LLM usage

LLM is involved in the production of this paper in the following ways:

1. Polish the human-written manuscripts, correcting grammar and spelling errors, enhancing readability and clarity of the paper.

[7] https://huggingface.co/facebook/hubert-large-ls960-ft

[8] https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification

[9] https://github.com/tarepan/SpeechMOS

2. Assisting in implementing the code for dataset preprocessing, as well as the collection and visualization of the evaluation results.