AffectCodec: Emotion-Preserving Neural Speech Codec for Expressive Speech Modeling

Jiacheng Shi[♠], **Hongfei Du**[♠], **Xinyuan Song**[♣], **Y. Alicia Hong**[♡]
Yanfu Zhang[♠], **Ye Gao**[♠]
[♠]College of William & Mary, [♣]Emory University, [♡]George Mason University
{jshi12, hdu02, yzhang105, ygao18}@wm.edu, xinyuan.song@emory.edu, yhong22@gmu.edu

Abstract

Neural speech codecs provide discrete representations for speech language models, but emotional cues are often degraded during quantization. Existing codecs mainly optimize acoustic reconstruction, leaving emotion expressiveness insufficiently modeled at the representation level. We propose an emotion-guided neural speech codec that explicitly preserves emotional information while maintaining semantic fidelity and prosodic naturalness. Our framework combines emotion–semantic guided latent modulation, relation-preserving emotional–semantic distillation, and emotion-weighted semantic alignment to retain emotionally salient cues under compression. Extensive evaluations across speech reconstruction, emotion recognition, and downstream text to speech generation demonstrate improved emotion consistency and perceptual quality without sacrificing content accuracy.

1 Introduction

Recent advances in large language models have rapidly extended to speech generation, enabling zero-shot text-to-speech (Wang et al., 2023), conversational voice agents (Zeng et al., 2024), and cross-modal audio reasoning (Shi et al., 2025b). A key enabler of this progress is the neural speech codec (Défossez et al., 2022; Zeghidour et al., 2021), which converts continuous waveforms into discrete representations. By transforming high-rate acoustic signals into compact symbol sequences at fixed frame rates, neural codecs bridge the gap between continuous speech and token-based sequence learners, making large-scale training and inference computationally tractable.

As neural codecs are increasingly adopted as the discrete representation layer for speech language models, their ability to preserve emotional information has emerged as a critical concern. Recent benchmark studies (Wu et al., 2024b) show that, despite strong performance in reconstructing linguistic content and speaker characteristics, modern neural codecs exhibit substantial variation in emotion preservation. Large-scale evaluations consistently report degradation in downstream emotion recognition when speech is resynthesized through codecs, with sensitivity to bitrate, architecture, and training data. More fine-grained analyses (Ren et al., 2024) further indicate that subtle and expressive emotional cues are particularly vulnerable to distortion, even in state-of-the-art neural codecs, leading to a measurable loss of emotion integrity and expressiveness relative to original speech. Notably, such degradation often occurs despite high overall reconstruction quality, suggesting that emotional information is more fragile than other speech attributes. Taken together, these findings imply that existing neural codecs largely preserve emotion as an implicit byproduct of compression, rather than explicitly modeling emotional expressiveness or its interaction with prosody and semantic content during representation learning. These observations naturally motivate the following research question: *How can a neural speech codec preserve emotion integrity and expressiveness while simultaneously maintaining prosodic naturalness and semantic fidelity under discrete representation?*

To address this question, we propose an emotion-guided neural speech codec that reconsiders the optimization priorities of discrete speech representation learning. Unlike prior neural codecs that primarily emphasize acoustic reconstruction (Défossez et al., 2022; Zeghidour et al., 2021) or semantic preservation (Ye et al., 2025a,b) and implicitly retain emotional information, our approach elevates emotion preservation to a primary modeling objective, while jointly maintaining prosodic naturalness and semantic fidelity. Our framework consists of three complementary stages. **(i) Emotion–Semantic Guided Latent Modulation** conditions acoustic latent representations on emotion- and

Findings of the Association for Computational Linguistics: ACL 2026, pages 9102–9124

semantics-related cues, enriching encoded features with affectively salient information prior to quantization. **(ii) Relation-Preserving Emotional–Semantic Distillation** encourages the codec to retain emotion-related relational structure during representation transformation, mitigating affective degradation introduced by discrete quantization. **(iii) Emotion-Weighted Semantic Alignment** further reinforces the association between discrete tokens and emotionally expressive content by adaptively weighting semantic alignment according to emotional salience. Together, these stages form a unified codec framework that explicitly prioritizes emotional expressiveness while preserving the structural constraints required for fluent prosody and accurate semantic content. Our contributions can be summarized as follows:

- **Conceptual Contribution.** We reconceptualize emotion preservation in neural speech codecs from a downstream evaluation concern to a core representation learning objective, explicitly treating emotional expressiveness as a primary optimization target rather than a post-hoc byproduct of acoustic reconstruction.
- **Methodological Contribution.** To the best of our knowledge, we present the first emotion-guided neural speech codec that addresses emotion degradation through a unified three-stage framework, integrating emotion–semantic guided latent, relation-preserving emotional–semantic distillation, and emotion-weighted semantic alignment.
- **Experimental Contribution.** Extensive experiments demonstrate that our approach improves reconstruction quality and representation effectiveness, and achieves strong emotion-related performance on the EMO-SUPERB and Codec-SUPERB benchmarks, as well as in zero-shot speech synthesis.

2 Related Work

We investigate prior work on Neural Speech Codecs and Discrete Audio Representation (Appendix A.1) and Emotion-Aware Speech Representation Learning (Appendix A.2), as these two research directions form the foundation for modeling discrete speech representations and preserving affective information in generative pipelines.

3 Methods

3.1 Representation Backbone and Guidance

As shown in Figure 1, discrete latent units provide a compact and temporally aligned representation for codec-based speech modeling following (Défossez et al., 2022; Xin et al., 2024b). Given an input waveform $\mathbf{x}$, we apply an acoustic encoder E_{a} to obtain a continuous latent sequence $\mathbf{A} = \{\mathbf{a}_t\}_{t=1}^{T'}$, where $\mathbf{a}_t \in \mathbb{R}^D$ denotes the frame-level representation and T' is the number of encoded frames. These latents capture fine-grained spectral and prosodic information and form the basis for subsequent emotion-guided refinement. To obtain discrete representations, we adopt a residual vector quantization (RVQ) module with K sequential codebook layers. At each layer k, the quantizer selects a codebook index sequence $\{q_t^{(k)}\}_{t=1}^{T'}$ and maps it to the corresponding embedding sequence $\mathbf{Q}^{(k)} \in \mathbb{R}^{T' \times D}$. The residual structure allows successive layers to model remaining quantization errors, yielding a refined approximation of the encoder latents. After K layers, we obtain the full discrete acoustic representation $\mathbf{Q}^{(1:K)}$. Table 1 reports a comparison with prior codecs.

Speech Emotion representation. We extract emotion-related features using a frozen emotion recognition model G_{emo}, which maps the input waveform $\mathbf{x}$ to a sequence of frame-level embeddings $\mathbf{E} = \{\mathbf{e}_t\}_{t=1}^{T'}$, where $\mathbf{e}_t \in \mathbb{R}^{D_{\mathrm{e}}}$. These embeddings provide emotion guidance for subsequent representation learning. When multiple hidden layers are available, we aggregate the layer-wise representations by averaging, $\mathbf{e}_t = \frac{1}{L_{\mathrm{e}}} \sum_{l=1}^{L_{\mathrm{e}}} \mathbf{h}_{\mathrm{emo},t}^{(l)}$, yielding a stable emotion embedding sequence $\mathbf{E}$ that serves as an emotion prior.

Audio Semantic representation. We extract semantic audio features using a frozen self-supervised speech model H. Given the input waveform $\mathbf{x}$, the model produces a sequence of frame-level hidden states, which we aggregate across layers to obtain a sequence of semantic audio embeddings $\mathbf{S} = \{\mathbf{s}_t\}_{t=1}^{T'}$, where $\mathbf{s}_t \in \mathbb{R}^{D_{\mathrm{s}}}$. Specifically, when multiple hidden layers are available, we compute each semantic embedding by layer averaging, $\mathbf{s}_t = \frac{1}{L} \sum_{\ell=1}^{L} \mathbf{h}_t^{(\ell)}$, yielding a stable semantic representation sequence $\mathbf{S}$ that provides high-level semantic guidance for subsequent modules.

Textual semantic representation. We extract textual semantic features using a frozen automatic speech recognition (ASR) model followed by a

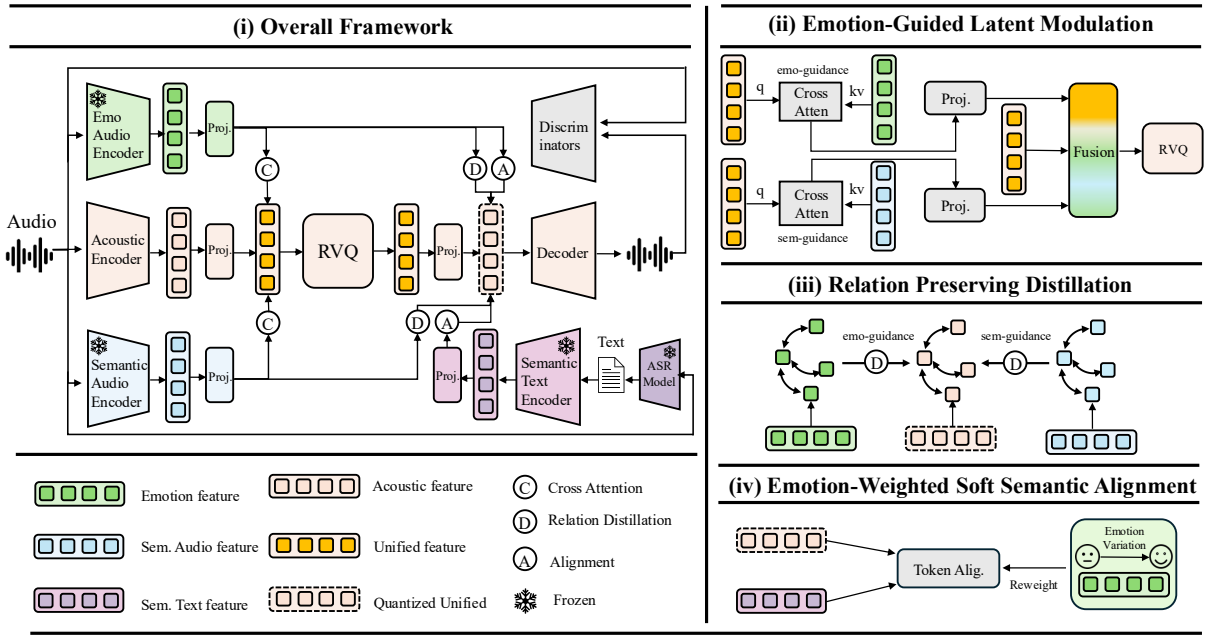

Figure 1: Overview of the proposed emotion-guided neural speech codec. The codec encodes input speech into discrete acoustic representations via residual vector quantization (RVQ) and incorporates emotion- and semantic-aware mechanisms to preserve emotional expressiveness. Specifically, it integrates (ii) emotion-guided latent modulation, which injects affective and semantic cues into acoustic latents prior to quantization, (iii) relation-preserving distillation, which constrains discrete representations to retain relational structure from emotion and semantic spaces, and (iv) emotion-weighted semantic alignment, which aligns quantized tokens with textual semantics while emphasizing emotionally salient regions to maintain semantic fidelity and prosodic naturalness.

pre-trained language encoder. Given the input waveform $\mathbf{x}$, the ASR model produces a token sequence, which is then processed by the language encoder to obtain a sequence of token-level embeddings $\mathbf{C} = \{\mathbf{c}_t\}_{t=1}^{T}$, where $\mathbf{c}_t \in \mathbb{R}^{D_c}$. When multiple transformer layers are available, we aggregate the layer-wise representations by averaging, $\mathbf{c}_t = \frac{1}{L}\sum_{\ell=1}^{L} \mathbf{h}_t^{(\ell)}$, yielding a stable textual semantic representation sequence $\mathbf{C}$ that provides linguistic guidance for subsequent modules.

3.2 Emotion-Guided Optimization

Standard neural speech codecs primarily optimize acoustic reconstruction (Défossez et al., 2022; Xin et al., 2024b) or semantic preservation (Ye et al., 2025a,b), leaving emotional information implicitly encoded and vulnerable to degradation during quantization. This indicates that existing representations lack explicitly mechanisms to model emotional expressiveness or its interaction with acoustic and semantic structure. To address this limitation, we propose an emotion-guided optimization framework with three stages: *Emotion-Semantic Guided Latent* (§3.2.1), which injects emotion- and semantic-aware signals into acoustic latents prior to quantization; *Relation-Preserving Emotional–Semantic Distillation* (§3.2.2), which preserves emotion-related relational structure; and *Emotion-Weighted Semantic Alignment* (§3.2.3), which strengthens the association between discrete tokens and emotionally expressive semantics.

3.2.1 Emotion–Semantic Guided Latent

Given frame-level acoustic latents $\mathbf{Z} = \{\mathbf{z}_t\}_{t=1}^{T'}$, we incorporate emotion and semantic guidance prior to quantization. Specifically, we extract emotion embeddings $\mathbf{E} = \{\mathbf{e}_t\}_{t=1}^{T'}$ using a frozen emotion encoder and semantic audio embeddings $\mathbf{S} = \{\mathbf{s}_t\}_{t=1}^{T'}$ using a pretrained self-supervised speech model. All representations are projected into a shared interaction space via linear transformations, yielding $\tilde{\mathbf{z}}_t = W_a\mathbf{z}_t$, $\tilde{\mathbf{e}}_t = W_e\mathbf{e}_t$, and $\tilde{\mathbf{s}}_t = W_s\mathbf{s}_t$, where $W_a \in \mathbb{R}^{D\times D}$, $W_e \in \mathbb{R}^{D\times D_e}$, and $W_s \in \mathbb{R}^{D\times D_s}$. We use acoustic features as queries in cross-attention over projected emotion and semantic sequences, where $\tilde{\mathbf{E}} = \{\tilde{\mathbf{e}}_t\}_{t=1}^{T'}$ and $\tilde{\mathbf{S}} = \{\tilde{\mathbf{s}}_t\}_{t=1}^{T'}$, yielding $\mathbf{h}_t^{\text{emo}} = \text{CrossAttn}(\tilde{\mathbf{z}}_t, \tilde{\mathbf{E}}, \tilde{\mathbf{E}})$ and $\mathbf{h}_t^{\text{sem}} = \text{CrossAttn}(\tilde{\mathbf{z}}_t, \tilde{\mathbf{S}}, \tilde{\mathbf{S}})$. The resulting modulation signals are projected back to the acoustic latent space using a shared linear mapping, producing $\mathbf{u}_t^{\text{emo}} = W_m\mathbf{h}_t^{\text{emo}}$ and $\mathbf{u}_t^{\text{sem}} = W_m\mathbf{h}_t^{\text{sem}}$. To balance emotion and semantic contributions, we apply independent stochastic dropout to each modulation term and form the unified latent representation as $\mathbf{z}_t^{\text{uni}} = \mathbf{z}_t + (\mathbf{u}_t^{\text{emo}} \odot \mathbf{d}_t^{\text{emo}}) + (\mathbf{u}_t^{\text{sem}} \odot \mathbf{d}_t^{\text{sem}})$,

where $\mathbf{d}_t^{\text{emo}}, \mathbf{d}_t^{\text{sem}} \in \{0,1\}^D$ are independently sampled. The resulting unified latents integrate emotion-aware and semantic-aware refinements while preserving the underlying acoustic structure prior to residual vector quantization.

3.2.2 Relation-Preserving Emotion–Semantic Distillation

Residual vector quantization (RVQ) can disrupt relational structure when mapping continuous representations to discrete codes. To address this issue, we adopt a relation-preserving distillation strategy (Wang et al., 2024b) that supervises the first-layer quantized representations by aligning their pairwise geometric relations with those from emotion and semantic teacher spaces. Given frame-level emotion features $\mathbf{E} = \{\mathbf{e}_t\}_{t=1}^{T'} \in \mathbb{R}^{T' \times D_{\text{e}}}$ and semantic audio features $\mathbf{S} = \{\mathbf{s}_t\}_{t=1}^{T'} \in \mathbb{R}^{T' \times D_{\text{s}}}$, we define teacher relational descriptors for each timestep pair (t, t') using Euclidean distances: $r_{t,t'}^{\text{emo}} = \|\mathbf{e}_t - \mathbf{e}_{t'}\|_2$ and $r_{t,t'}^{\text{sem}} = \|\mathbf{s}_t - \mathbf{s}_{t'}\|_2$. Prior work has shown that the first RVQ layer captures particularly informative and structured representations (Xin et al., 2024b). Accordingly, we compute student relational descriptors from the first residual quantized outputs $\mathbf{Q}^{(1)} = \{\mathbf{Q}_t^{(1)}\}_{t=1}^{T'}$ as $r_{t,t'}^{(1)} = \|\mathbf{Q}_t^{(1)} - \mathbf{Q}_{t'}^{(1)}\|_2$. Relational consistency between student and teacher is enforced via

$$\mathcal{L}_{\text{rela}} = \frac{1}{T'^2} \sum_{t,t'} \left[\alpha \, d\Big(r_{t,t'}^{(1)}, r_{t,t'}^{\text{emo}}\Big) + \beta \, d\Big(r_{t,t'}^{(1)}, r_{t,t'}^{\text{sem}}\Big) \right], \tag{1}$$

where $d(\cdot,\cdot)$ denotes an ℓ_1 discrepancy. This objective preserves emotion- and semantic-relevant relational structure under discretization.

3.2.3 Emotion-Weighted Semantic Alignment

To address the inherent mismatch in sequence length between frame-level RVQ outputs and token-level contextual embeddings, we perform semantic alignment between quantized representations and textual semantics. Moreover, to mitigate emotion degradation introduced by discrete quantization, we impose stronger supervision on frames exhibiting larger emotion variation, which are empirically more susceptible to affective distortion. The complete procedure is summarized in Algorithm 1.

Let the quantized latent sequence be $\mathbf{Q}^{(1)} = \{\mathbf{Q}_t^{(1)}\}_{t=1}^{T'}$, the textual semantic embeddings be $\mathbf{C} = \{\mathbf{c}_i\}_{i=1}^{n}$, and the frame-level emotion features be $\mathbf{E} = \{\mathbf{e}_t\}_{t=1}^{T'}$. Assuming a roughly monotonic correspondence between speech frames and text tokens, each frame t is associated with a local textual neighborhood without dynamic programming. We compute a center index $i_0(t) = \lfloor tn/T' \rfloor$ and define a window $\mathcal{N}(t)$ of width $2w+1$. Within this window, cosine similarities between $\mathbf{Q}_t^{(1)}$ and $\mathbf{c}_i$ are normalized via softmax to obtain alignment weights $a_{t,i}$, which are used to construct a semantic teacher $\mathbf{c}_t^* = \sum_{i \in \mathcal{N}(t)} a_{t,i} \mathbf{c}_i$. To explicitly realize this emotion-aware supervision, *Emotion Variation* is operationalized as the magnitude of frame-level emotion variation, quantified by differences in emotion embeddings across adjacent frames. Accordingly, we compute a framewise emotion difference magnitude as $d_t = \|\mathbf{e}_t - \mathbf{e}_{t-1}\|_1$ with $d_1 = 0$, and derive normalized importance weights γ_t by applying a softmax over $\{d_t\}$ and scaling to unit mean. These weights modulate each frame's contribution in the alignment objective, which is defined as

$$\mathcal{L}_{\text{align}} = -\frac{1}{T'} \sum_{t=1}^{T'} \gamma_t \log \sigma\Big(\cos(\mathbf{Q}_t^{(1)}, \mathbf{c}_t^*)\Big), \tag{2}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

Algorithm 1 Emo-Weighted Semantic Alignment

Require: Quantized representations $\{\mathbf{Q}_t^{(1)}\}_{t=1}^{T'}$, textual semantic embeddings $\{\mathbf{c}_i\}_{i=1}^{n}$, emotion features $\{\mathbf{e}_t\}_{t=1}^{T'}$, window size w

Compute framewise emotion differences:
$d_1 = 0, \quad d_t = \|\mathbf{e}_t - \mathbf{e}_{t-1}\|_1, \quad t = 2, \ldots, T'$
Compute emotion importance weights:
$\boldsymbol{\gamma} = T' \cdot \text{Softmax}(\{d_t\}_{t=1}^{T'})$
for $t = 1$ to T' **do**
  Compute center index:
  $i_0 = \text{clip}\big(\lfloor t \cdot \frac{n}{T'} \rfloor, \ 1, \ n\big)$
  Define local neighborhood:
  $\mathcal{N}(t) = \{\, i \mid |i - i_0| \le w \,\}$
  Compute cosine similarities:
  $s_{t,i} = \cos(\mathbf{Q}_t^{(1)}, \mathbf{c}_i), \quad i \in \mathcal{N}(t)$
  Compute alignment weights:
  $a_{t,i} = \exp(s_{t,i}) / \sum_{j \in \mathcal{N}(t)} \exp(s_{t,j})$
  Construct semantic teacher:
  $\mathbf{c}_t^* = \sum_{i \in \mathcal{N}(t)} a_{t,i} \, \mathbf{c}_i$
end for
return $\{\mathbf{c}_t^*\}_{t=1}^{T'}$ and $\{\gamma_t\}_{t=1}^{T'}$

3.3 Training Objective

We adopt a multi-objective training framework that combines standard neural codec reconstruction losses (Défossez et al., 2022; Xin et al., 2024b;

Ji et al., 2025) with two emotion–semantic supervisory objectives. The generator is optimized with four reconstruction losses: the quantization commitment loss $\mathcal{L}_q$, the mel-spectrogram loss $\mathcal{L}_{\text{mel}}$, the adversarial loss $\mathcal{L}_{\text{adv}}$, and the feature matching loss $\mathcal{L}_{\text{feat}}$. In addition, we regularize discrete representations using two objectives: the relation-preserving distillation loss $\mathcal{L}_{\text{rela}}$, which enforces emotion and semantic relational consistency from teacher representations, and the emotion-weighted semantic alignment loss $\mathcal{L}_{\text{align}}$, which aligns quantized tokens with contextual semantics while emphasizing emotionally salient frames. The overall training objective is defined as

$$\mathcal{L}_{\text{total}} = \lambda_{\text{mel}}\mathcal{L}_{\text{mel}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} + \lambda_q\mathcal{L}_q + \lambda_{\text{rela}}\mathcal{L}_{\text{rela}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}. \quad (3)$$

Detailed training objectives are in Appendix G.4.

3.4 Downstream Extension to TTS

The emotion-aware discrete representations learned by the codec are further utilized for text-to-speech synthesis by training a language model over RVQ tokens, following prior work (Xin et al., 2024b; Wang et al., 2023). Conditioned on a phoneme sequence $\mathbf{y}$ and a reference acoustic prompt $\mathbf{P} \in \mathbb{R}^{T' \times K}$, the model generates K parallel token streams aligned with the RVQ hierarchy. The first stream, which encodes coarse linguistic structure and global prosodic, is modeled autoregressively using a decoder-only Transformer, optimized with

$$\mathcal{L}_{\text{AR}} = -\log \prod_{t=1}^{T'} p\Big(u_t^{(1)} \;\Big|\; u_{<t}^{(1)}, \mathbf{y}; \theta_{\text{AR}}\Big). \quad (4)$$

Subsequent RVQ streams encode finer acoustic detail and emotional variation. For layers $k = 2, \ldots, K$, a non-autoregressive Transformer infers the entire token sequence $\mathbf{u}^{(k)}$ conditioned on previously inferred layers, the phoneme sequence, and the acoustic prompt:

$$\mathcal{L}_{\text{NAR}} = -\log \prod_{k=2}^{K} p\Big(\mathbf{u}^{(k)} \;\Big|\; \mathbf{u}^{(<k)}, \mathbf{y}, \mathbf{P}; \theta_{\text{NAR}}\Big). \quad (5)$$

Both components share an identical Transformer configuration with 12 layers, 16 attention heads, 1024-dimensional embeddings, 4096-dimensional feed-forward layers, and a dropout rate of 0.05. The inferred RVQ token hierarchies are subsequently mapped to discrete embeddings $\mathbf{Q}^{(k)}$ and decoded to synthesize speech under joint textual, acoustic, and emotion-aware conditioning.

Model	BR↓	FR↓	Nq↓	Train. Data↓	Param↓
Encodec	6	75	8	17	20
DAC	6	50	12	8	76
FACodec	4.8	80	6	500	500
BigCodec	1	80	1	1	159
Mimi	1.1	12.5	8	1	95
TAAE	0.6	25	1	100	950
WavTokenizer	0.9	75	1	80	40
SpeechTokenizer	4	50	8	1	18
Llasa	0.8	50	1	150	1000
Ours	4	50	8	2.3	44

Table 1: Comparison of codecs across efficiency dimensions. BR: Bitrate (kbps); FR: Frame Rate (Hz); Nq: Number of Quantizers; Train. Data: Trainng Hours (k hours); Param: Number of Parameters (M).

4 Experiments

4.1 Experiment Setups

Datasets. We train our model on a multi-domain speech corpus of approximately 2.3K hours, covering three complementary aspects: clean read speech for high-fidelity reconstruction, multilingual and acoustically diverse speech for robustness, and emotionally expressive speech for affective representation learning. Specifically, LibriSpeech (Panayotov et al., 2015) and VCTK (Yamagishi et al., 2019) are used as clean English read-speech sources for reconstruction, AISHELL-3 (Shi et al., 2020) provides Mandarin speech for multilingual generalization, and a subset of AudioSet (Gemmeke et al., 2017) introduces broad acoustic variability. To expose the codec to natural emotional expressions, we additionally incorporate emotionally annotated conversational speech from MSP-Podcast (Busso et al., 2025) and CMU-MOSEI (Bagher Zadeh et al., 2018). All audio is resampled to 16 kHz. For evaluation, speech reconstruction quality is measured on the LibriSpeech test-clean set (Panayotov et al., 2015). To assess emotion preservation under codec resynthesis, we follow the EMO-SUPERB (Wu et al., 2024a) protocol and evaluate speech emotion recognition on synthesized speech using the official EMO-SUPERB test partitions. For downstream text-to-speech generation, the token prediction module is trained on LibriTTS (Zen et al., 2019), and evaluation covers linguistic intelligibility on LibriSpeech test-clean and emotion preservation on EmoVoiceDB (Wu et al., 2024a) and SECAP (Xu et al., 2024). Detailed dataset are in the Appendix C.

Architecture. The speech tokenizer comprises an acoustic encoder, a residual vector quantization (RVQ) module with eight quantization layers and codebooks of size 1024, and a decoder $\text{Dec}(\cdot)$,

Table 2: **Objective speech reconstruction results** are reported across emotional consistency, information preservation, and speech naturalness, evaluated on EmoVoiceDB and the LibriSpeech test-clean set. **Bold** marks best scores, and underline indicates second-best scores. Results are averaged over three random seeds.

Model	Emotional Consistency			Information Preservation			Speech Naturalness		
	Emo SIM↑	Pros SIM↑	Recall↑	WER↓	WIL↓	LSD↓	MSEP↓	PESQ↑	UTMOS↑
EnCodec	0.73	0.78	0.37	**4.02**	6.63	0.97	35.06	2.38	2.43
DAC	0.79	0.74	0.31	4.10	6.52	0.94	30.36	2.74	3.31
FACodec	0.88	0.70	0.32	4.14	6.64	0.85	**12.16**	2.85	3.49
SpeechTokenizer	0.82	0.77	0.29	4.19	6.78	1.07	40.38	2.58	3.43
Mimi	0.85	0.78	0.35	10.72	16.34	1.08	43.24	1.65	2.29
BigCodec	0.78	0.71	0.32	4.58	7.45	0.84	22.91	2.68	3.44
TAAE	0.84	0.73	0.33	9.35	13.81	1.05	42.35	2.03	3.57
WavTokenizer	0.83	0.81	0.38	6.22	9.16	0.96	39.35	2.19	3.36
Llasa	0.87	0.80	0.40	4.46	7.20	0.92	27.49	2.43	3.55
Ours	**0.94**	**0.86**	**0.48**	4.15	**6.43**	**0.78**	19.21	**3.04**	**3.68**

following the setting in (Xin et al., 2024b). Adversarial training employs three discriminators, including multi-period, multi-scale, and multi-scale STFT discriminators. Quantization is performed at 50 Hz, with both the encoder and RVQ using an embedding dimension of $D = 1024$. To provide affective and semantic guidance, we incorporate several frozen pre-trained encoders, including CLAP-LAION (630k-best) (Wu et al., 2023b) as the emotion encoder, wav2vec 2.0 (base-960h) (Baevski et al., 2020) as the ASR model, BERT (bert-base) (Devlin et al., 2019) as the language encoder, and HuBERT (base-ls960) (Hsu et al., 2021) as the self-supervised speech model. All pre-trained encoders output embeddings with $D_e = D_s = D_c = 768$. Cross-attention modules are implemented with eight attention heads. Additional details are in the Appendix G.

Implementation Details and metrics. Our codec model is trained for 200 epochs on four A100 GPUs with a batch size of 16, using the AdamW optimizer with a learning rate of 2×10^{-4} and the learning rate is decayed based on a cosine schedule. We set $\alpha = \beta = 1$ for distillation. The downstream TTS models are trained on four A100 GPUs using ScaledAdam with a learning rate of 5×10^{-2} and 120 warm-up steps, where the AR model is trained for 300 epochs and the NAR model for 200 epochs. Training employs dynamic batching, with each batch containing up to 550 seconds of audio for AR and 100–200 seconds for NAR.

We evaluate our model along three dimensions: emotional consistency, content preservation, and speech naturalness. Emotional consistency is assessed using Emotion Similarity computed from emotion2vec embeddings (Ma et al., 2024), Prosody Similarity derived from pitch, energy, and duration features via AutoPCP (Barrault et al., 2023; Wu et al., 2024c), and emotion recognition recall measured by a pretrained SER model (Ma et al., 2024). Content preservation is evaluated through Whisper-based transcription accuracy, including Word Error Rate (WER), Word Information Lost (WIL)(Morris et al., 2004), and Log-Spectral Distance (LSD). Speech naturalness is measured using pitch reconstruction error (MSEP), PESQ (Rix et al., 2002) for perceptual quality under signal distortion, and UTMOS (Chen et al., 2022), which predicts human-judged speech naturalness.

4.2 Main Results

4.2.1 Evaluation on Reconstruction

We evaluate reconstruction performance on two benchmarks, as shown in Table 2: EmoVoiceDB, which assesses emotion preservation in reconstructed speech, and LibriSpeech test-clean, which measures intelligibility and perceptual quality following prior work (Ji et al., 2024; Xin et al., 2024a). Complete baseline details are in the Appendix B.

Emotional Consistency. Our method achieves the strongest emotion preservation across all metrics, with Emo SIM of 0.94 surpassing FACodec (0.88) and Pros SIM of 0.86 exceeding WavTokenizer (Ji et al., 2025) (0.81). SER-based emotion recall reaches 0.48, notably higher than Llasa (0.40), indicating improved preservation of both global emotion and fine-grained prosody.

Information Preservation. On LibriSpeech, the proposed codec maintains competitive intelligibility while explicitly optimizing emotion preservation. Although EnCodec achieves the lowest WER (4.02), our method remains comparable (4.15) and

Table 3: **Speech emotion recognition performance** on the EMO-SUPERB benchmark. Macro-F1 scores are reported for six emotion-focused evaluation sets. **Bold** indicates the best result, and underline marks the second-best. Results are averaged over three random seeds.

Model	Codec Information		Speech Emotion Recognition (Macro-F1) ↑					
	kbps	Configuration	IEMOCAP	CREMA-D	IMPROV	PODCAST	NNIME	BIIC-POD.
Original Audio	-	-	0.313	0.594	0.491	0.301	0.183	0.247
AudioDec	6.4	symAD_libritts_24000_hop300	0.301	0.548	0.461	0.298	0.180	0.242
AcademiCodec	2	large universal	0.301	0.548	0.461	0.298	0.181	0.242
SpeechTokenizer	4	hubert_avg	0.305	0.573	0.448	0.292	0.180	0.243
DAC	6	DAC_16k	0.315	0.591	0.491	0.302	**0.184**	0.247
EnCodec	1.5	24k	0.280	0.411	0.321	0.262	0.166	0.227
EnCodec	3	24k	0.275	0.457	0.448	0.293	0.178	0.240
EnCodec	6	24k	0.295	0.499	0.450	0.294	0.178	0.239
FunCodec	8	en_libritts_16k_nq32ds640	0.312	0.569	0.482	0.303	0.181	0.246
FunCodec	8	zh_en_16k_nq32ds640	0.312	0.577	0.482	0.302	0.182	0.246
Soundstream	6	Soundstream	0.261	0.411	0.321	0.262	0.146	0.213
MP3	6	-	0.259	0.405	0.321	0.262	0.149	0.210
Opus	6	-	0.271	0.433	0.338	0.269	0.162	0.226
AAC	6	-	0.268	0.428	0.334	0.268	0.165	0.227
Ours	4	16k	**0.338**	**0.629**	**0.513**	**0.319**	0.182	**0.256**

yields stronger complementary content metrics, including lower WIL (6.43) and the lowest LSD (0.78). These results indicate that emotion-aware modeling preserves spectral detail and linguistic content without sacrificing expressiveness.

Speech Naturalness. Our codec achieves the best perceptual quality, with the highest PESQ (3.04) and UTMOS (3.68). Although MSEP is not the lowest, improvements in perceptual metrics indicate cleaner and more natural reconstructions. Our codec better balances emotional fidelity, content preservation, and perceptual naturalness than prior approaches. Subjective results and representation quality evaluation are in the Appendix D and E.

4.2.2 Speech Emotion Recognition Evaluation

Reconstruction-based metrics do not directly reflect whether emotion-related information remains usable for downstream perception. We therefore evaluate our codec on the EMO-SUPERB benchmark, which measures emotion discrimination performance across six standard SER datasets using macro-F1. Results are reported in Table 3.

Trend across English datasets. Across IEMOCAP (Busso et al., 2008), CREMA-D (Cao et al., 2014), and IMPRoV (Busso et al., 2016), our model consistently ranks first or second among all codec baselines, indicating strong emotion retention. Notably, on IMPRoV, our representation even outperforms original audio, suggesting that the learned discrete space suppresses nuisance variability such as channel mismatch and recording artifacts, which benefits downstream emotion classification.

Variability on Chinese datasets. On BIIC-PODCAST (Upadhyay et al., 2023), our codec achieves the best F1 score, showing robustness to conversational speech and diverse recording conditions. Performance on NNIME (Chou et al., 2017) is slightly below the strongest baseline, due to its subtle and regulated emotional expressions, which remain challenging for discrete representations.

General observations. Our codec generalizes across languages, outperforming neural and legacy codecs at comparable bitrates and approaching original emotion discriminability.

4.2.3 Zero-shot TTS Generation Evaluation

To assess whether the learned codec representations generalize beyond reconstruction, we evaluate their effectiveness in zero-shot text-to-speech (TTS) synthesis (Table 4). Following prior work, the TTS model is trained on LibriTTS and evaluated on three benchmarks: LibriSpeech for intelligibility and prosodic naturalness (WER, WIL, SIM-O (Ye et al., 2025a), UTMOS), and EmoVoice-DB and SECAP for emotion alignment (Emo SIM, Recall) under expressive speech. Our goal is not to optimize TTS quality itself, but to examine whether the codec representations support intelligible, prosodically coherent, and emotionally aligned synthesis. Subjective results are provided in Appendix E.

Linguistic precision and prosodic fluency. On LibriSpeech, our codec achieves the highest prosody similarity (SIM-O = 0.80) with competitive naturalness (UTMOS = 4.29) and strong intelligibility (WER = 2.51). Although CosyVoice-2 (Du et al., 2024a) attains slightly lower WER, it shows weaker prosodic and perceptual quality. On the emotion-rich EmoVoice-DB and SECAP datasets, our method achieves the highest UTMOS scores, indicating stable expressive pitch and rhythm.

Emotion discrimination during synthesis. Our

Table 4: **TTS Evaluation Results.** We evaluate zero-shot TTS performance across three datasets: LibriSpeech, EmoVoiceDB, and SECAP. Each evaluation reports content preservation (WER), speaker similarity (SIM), emotional alignment (Emo_SIM), and naturalness (UTMOS). Results are averaged over three random seeds.

System	Frame Rate	LibriSpeech				EmoVoice-DB				SECAP			
		WER↓	SIM-O↑	WIL↓	UTMOS↑	WER↓	Emo_SIM↑	Recall↑	UTMOS↑	WER↓	Emo_SIM↑	Recall↑	UTMOS↑
NAR Models													
MaskGCT	50	2.63	0.68	13.83	3.10	3.28	0.71	0.33	3.25	9.52	0.74	0.36	2.70
F5-TTS	93.75	2.53	0.66	11.10	3.25	3.45	0.69	0.31	3.30	8.87	0.72	0.34	2.65
AR Models													
FireRedTTS	25	2.69	0.47	15.47	3.05	3.61	0.57	0.29	3.10	9.13	0.59	0.31	2.75
ARS	50	2.64	0.68	10.88	3.45	3.42	0.74	0.34	3.50	8.65	0.72	0.38	2.95
CosyVoice 2	25	**2.45**	0.77	**6.72**	4.23	3.47	0.87	0.37	**4.42**	**8.55**	0.79	0.43	2.75
Llasa	50	2.49	0.58	9.34	3.55	3.61	0.70	0.32	3.40	8.92	0.69	0.35	2.85
SparkTTS	50	2.57	0.78	7.18	4.17	**3.28**	0.82	0.36	4.20	9.03	0.77	0.40	2.95
Ours	50	2.51	**0.80**	7.29	**4.29**	3.48	**0.91**	**0.41**	4.35	8.64	**0.84**	**0.49**	**3.20**

Model Components			Reconstruction				TTS			
EG-Latent	RP-Distill	EW-Align	Emo SIM ↑	WER ↓	UTMOS ↑	MUSHRA ↑	Emo SIM ↑	WER ↓	UTMOS ↑	MOS ↑
			0.87	4.85	3.34	86.27	0.86	4.12	3.85	3.68
✓			0.90	4.62	3.46	88.31	0.88	3.95	3.98	3.80
	✓		0.89	4.47	3.41	87.54	0.89	3.69	4.01	3.77
		✓	0.88	4.53	3.49	87.12	0.87	3.81	4.06	3.83
✓	✓		0.93	4.30	3.54	89.63	0.90	3.55	4.15	3.92
✓		✓	0.92	4.42	3.57	89.94	0.89	3.62	4.22	3.98
	✓	✓	0.90	4.21	3.52	88.86	0.90	3.51	4.18	3.90
✓	✓	✓	**0.94**	**4.15**	**3.68**	**90.71**	**0.91**	**3.48**	**4.35**	**4.05**

Table 5: Ablation study on the three components: **E**motion-**S**emantic **G**uided **Latent** (EG-Latent), **R**elation-**P**reserving **Distill**ation (RP-Distill), and **E**motion-**W**eighted Soft Semantic **Align**ment (EW-Align). We assess emotional expressiveness (Emo SIM), semantic preservation (WER), prosodic naturalness (UTMOS), and subjective perceptual quality (MUSHRA/MOS) for both reconstruction and TTS. Following prior settings, LibriSpeech evaluates semantic fidelity and prosodic naturalness, while EmoVoice-DB assesses emotion preservation.

model also preserves emotional intent during synthesis. It achieves the highest Emo_SIM and recall on both EmoVoice-DB and SECAP, confirming that the emotion-aware cues encoded in the latent space remain discriminative after generation. Overall, these results show that our codec enables zero-shot speech synthesis with improved affective realism and robustness in emotionally dynamic conditions. We release our demo publicly available.[1]

4.3 Ablation

Effect of EG-Latent. As shown in Table 5, introducing EG-Latent substantially enhances emotional expressiveness and reconstruction fidelity. Specifically, Emo SIM increases from 0.87 to 0.90, while MUSHRA (Défossez et al., 2022) improves from 86.27 to 88.31. These indicate that emotion-guided latent modeling enables the codec to better preserve affective nuances during reconstruction.

Effect of RP-Distill. RP-Distill aligns representations with semantic knowledge from a teacher, primarily improving linguistic clarity. Compared to the baseline, it reduces WER from 5.75 to 5.18 and slightly increases TTS Emo SIM from 0.86 to 0.89, indicating improved intelligibility while preserving emotional content for downstream generation.

Effect of EW-Align. EW-Align strengthens prosodic and semantic expression, particularly at emotionally salient regions. It increases reconstruction UTMOS from 3.34 to 3.49 and TTS Emo SIM from 0.86 to 0.87, highlighting gains in naturalness and emotion coherence. Overall, the components provide complementary improvements across emotion, content, and prosody, with their combination achieving the best performance across all metrics. Qualitative analysis and fine-grained ablations are provided in Appendix H and Appendix F.

5 Conclusion

We present AffectCodec, a neural speech codec that explicitly targets emotion preservation in discrete speech representations. By integrating emotion-aware latent, relation-preserving distillation, and emotion-weighted semantic alignment, the proposed codec improves emotional expressiveness while retaining semantic fidelity and prosodic naturalness. Extensive evaluations on speech reconstruction, speech emotion recognition, and zero-shot TTS synthesis demonstrate consistent gains in emotional consistency and perceptual quality across diverse speaking styles and conditions.

[1] https://jiachengqaq.github.io/affectcodec_demo/

6 Limitations

The proposed codec achieves strong performance on emotion-related speech reconstruction benchmarks, demonstrating effective preservation of affective cues alongside semantic and prosodic fidelity. The framework is designed to reconstruct emotional expressiveness under acceptable computational efficiency, rather than to minimize model complexity. Future work may explore lighter-weight architectures and more efficient training strategies to further improve scalability while retaining emotion-preserving capabilities.

7 Ethics Statement

All speech datasets used in this work, as detailed in Appendix C, are publicly released for academic research purposes. We strictly adhere to the licenses and usage terms associated with each dataset. The data do not contain personally identifiable information (PII), and no attempt is made to identify or infer individual identities from the speech signals. The proposed method is developed and evaluated solely in an offline research setting. No real-world deployment or user-facing application is involved in this work.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, and 1 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Marina Bosi, Karlheinz Brandenburg, Schuyler Quackenbush, Louis Fielder, Kenzo Akagiri, Hendrik Fuchs, and Martin Dietz. 1997. Iso/iec mpeg-2 advanced audio coding. *Journal of the Audio engineering society*, 45(10):789–814.

Karlheinz Brandenburg. 1994. Iso-mpeg-1 audio: A generic standard for coding of high-quality digital audio. *J. Audio Eng. Soc.*, 42(10):780–792.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Carlos Busso, Reza Lotfian, Kusha Sridhar, Ali N Salman, Wei-Cheng Lin, Lucas Goncalves, Srinivas Parthasarathy, Abinay Reddy Naini, Seong-Gyun Leem, Luz Martinez-Lucas, and 1 others. 2025. The msp-podcast corpus. *arXiv preprint arXiv:2509.09791*.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271.

Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee. 2017. Nnime: The nthu-ntua chinese interactive multimodal emotion corpus. In *2017 Seventh international conference on affective computing and intelligent interaction (ACII)*, pages 292–298. IEEE.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2015. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 4(5):11.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024a. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Zhihao Du, Shiliang Zhang, Kai Hu, and Siqi Zheng. 2024b. Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 591–595. IEEE.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen. 2025. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Hao-Han Guo, Yao Hu, Kun Liu, Fei-Yu Shen, Xu Tang, Yi-Chen Wu, Feng-Long Xie, Kun Xie, and Kai-Tuo Xu. 2024. Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications. *arXiv preprint arXiv:2409.03283*.

Yiwei Guo, Chenpeng Du, Xie Chen, and Kai Yu. 2023. Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE.

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

Shengpeng Ji, Ziyue Jiang, Hanting Wang, Jialong Zuo, and Zhou Zhao. 2024. Mobilespeech: A fast and high-fidelity framework for mobile zero-shot text-to-speech. *arXiv preprint arXiv:2402.09378*.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, and 1 others. 2025. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, and 1 others. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

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

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993.

Reza Lotfian and Carlos Busso. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760.

Andrew Cameron Morris, Viktoria Maier, and Phil D Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech*, pages 2765–2768.

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

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

Tianyi Peng and Yang Xiao. 2025. Dark experience for incremental keyword spotting. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Wenze Ren, Yi-Cheng Lin, Huang-Cheng Chou, Haibin Wu, Yi-Chiao Wu, Chi-Chun Lee, Hung-yi Lee, Hsin-Min Wang, and Yu Tsao. 2024. Emo-codec: An in-depth look at emotion preservation capacity of legacy and neural codec models with subjective and objective evaluations. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–6. IEEE.

Antony W Rix, Michael P Hollier, Andries P Hekstra, and John G Beerends. 2002. Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part i–time-delay compensation. *Journal of the Audio Engineering Society*, 50(10):755–764.

Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The interspeech 2009 emotion challenge.

Roneel V Sharan, Cecilia Mascolo, and Björn W Schuller. 2024. Emotion recognition from speech signals by mel-spectrogram and a cnn-rnn. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4. IEEE.

Jiacheng Shi, Hongfei Du, Yangfan He, Y Alicia Hong, and Ye Gao. 2025a. Emotion-aligned generation in diffusion text to speech models via preference-guided optimization. *arXiv preprint arXiv:2509.25416*.

Jiacheng Shi, Hongfei Du, Y Alicia Hong, and Ye Gao. 2025b. Plug-and-play emotion graphs for compositional prompting in zero-shot speech emotion recognition. *arXiv preprint arXiv:2509.25458*.

Jiacheng Shi, Yanfu Zhang, and Ye Gao. 2025c. Clepdg: Contrastive learning for speech emotion domain generalization via soft prompt tuning. *arXiv preprint arXiv:2507.04048*.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*.

Premjeet Singh, Shefali Waldekar, Md Sahidullah, and Goutam Saha. 2022. Analysis of constant-q filterbank based representations for speech emotion recognition. *Digital Signal Processing*, 130:103712.

Shreya G Upadhyay, Woan-Shiuan Chien, Bo-Hao Su, Lucas Goncalves, Ya-Tse Wu, Ali N Salman, Carlos Busso, and Chi-Chun Lee. 2023. An intelligent infrastructure toward large scale naturalistic affective speech corpora collection. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Jean-Marc Valin, Gregory Maxwell, Timothy B Terriberry, and Koen Vos. 2016. High-quality, low-delay music coding in the opus codec. *arXiv preprint arXiv:1602.04845*.

Aaron Van Den Oord, Oriol Vinyals, and 1 others. 2017a. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Aaron Van Den Oord, Oriol Vinyals, and 1 others. 2017b. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Dimitrios Ververidis and Constantine Kotropoulos. 2006. Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9):1162–1181.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. 2024a. Blsp-emo: Towards empathetic large speech-language models. *arXiv preprint arXiv:2406.03872*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Sijie Wang, Rui She, Qiyu Kang, Xingchao Jian, Kai Zhao, Yang Song, and Wee Peng Tay. 2024b. Distilvpr: Cross-modal knowledge distillation for visual place recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 10377–10385.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, and 1 others. 2025. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2024c. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*.

Haibin Wu, Huang-Cheng Chou, Kai-Wei Chang, Lucas Goncalves, Jiawei Du, Jyh-Shing Roger Jang, Chi-Chun Lee, and Hung-Yi Lee. 2024a. Emo-superb: An in-depth look at speech emotion recognition. *arXiv preprint arXiv:2402.13018*.

Haibin Wu, Ho-Lam Chung, Yi-Cheng Lin, Yuan-Kuei Wu, Xuanjun Chen, Yu-Chi Pai, Hsiu-Hsuan Wang, Kai-Wei Chang, Alex Liu, and Hung-yi Lee. 2024b. Codec-superb: An in-depth analysis of sound codec models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10330–10348.

Haibin Wu, Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Daniel Tompkins, Chung-Hsien Tsai, Canrun Li, Zhen Xiao, Sheng Zhao, Jinyu Li, and

1 others. 2024c. Laugh now cry later: Controlling time-varying emotional states of flow-matching-based zero-shot text-to-speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 690–697. IEEE.

Pengfei Wu, Zhenhua Ling, Lijuan Liu, Yuan Jiang, Hongchuan Wu, and Lirong Dai. 2019. End-to-end emotional speech synthesis using style tokens and semi-supervised training. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE.

Yi-Chiao Wu, Israel D Gebru, Dejan Marković, and Alexander Richard. 2023a. Audiodec: An open-source streaming high-fidelity neural audio codec. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023b. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yang Xiao and Rohan Kumar Das. 2024. Where's that voice coming? Continual learning for sound source localization. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*.

Yang Xiao and Rohan Kumar Das. 2025. UCIL: An unsupervised class incremental learning approach for sound event detection. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Yang Xiao, Han Yin, Jisheng Bai, and Rohan Kumar Das. 2025. Dg-sed: Domain generalization for sound event detection with heterogeneous training data. In *2025 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

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

Z Xin, Z Dong, L Shimin, Z Yaqian, and Q Xipeng. 2024b. Speechtokenizer: Unified speech tokenizer for speech language models. In *Proc. Int. Conf. Learn. Representations*, pages 1–21.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19323–19331.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive:(http://web. ku. edu/˜ idea/readings/rainbow. htm)*.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. 2023. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*.

Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, and 1 others. 2025. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10748–10757.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, and 1 others. 2025a. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25697–25705.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, and 1 others. 2025b. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*.

Han Yin, Yang Xiao, Rohan Kumar Das, Jisheng Bai, Haohe Liu, Wenwu Wang, and Mark D Plumbley. 2025. EnvSDD: Benchmarking environmental sound deepfake detection. In *Proc. Interspeech*, pages 201–205.

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

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.

Appendix

A Related Work

A.1 Neural Speech Codecs and Discrete Audio Representation

Neural audio codecs based on vector quantization (Van Den Oord et al., 2017a) have become a cornerstone of modern speech and audio generation systems. Early work such as VQ-VAE (Van Den Oord et al., 2017b) and its residual variants introduced discrete latent representations for efficient compression and autoregressive modeling. Building on this paradigm, SoundStream (Zeghidour et al., 2021) and EnCodec (Défossez et al., 2022) employed residual vector quantization with adversarial training to achieve high-fidelity reconstruction at low bitrates, while HiFi-Codec (Yang et al., 2023) and DAC (Kumar et al., 2023) further improved efficiency and stability through group quantization and refined codebook learning. These codecs primarily optimize acoustic reconstruction quality and compression efficiency, treating emotion as an implicit attribute of the signal.

More recent work has explored codec designs tailored for language-model-based speech generation. SpeechTokenizer (Xin et al., 2024b), FACodec (Ju et al., 2024), and Mimi (Défossez et al., 2024) introduce hierarchical or disentangled representations, often supervising semantic information in early quantization layers using ASR or self-supervised speech models. Single-codebook designs such as BigCodec (Xin et al., 2024a), WavTokenizer (Ji et al., 2025), and TAAE (Parker et al., 2024) improve compatibility with large language models by flattening token streams, but may limit expressiveness at low token rates. Semantic-enhanced codecs such as X-Codec (Ye et al., 2025a) incorporate pretrained speech representations into the quantization process to improve downstream modeling, while Llasa (Ye et al., 2025b) aligns speech tokenization with standard LLM architectures using a simplified codec and Transformer framework. Despite these advances, emotional expressiveness is typically not an explicit optimization target and remains weakly constrained.

EmoCodec (Ren et al., 2024) highlights this limitation by systematically evaluating emotional degradation in existing codecs, showing that affective cues are often poorly preserved after quantization (Wu et al., 2024b). However, emotion is primarily treated as an evaluation dimension rather than a modeling objective. In contrast, our work explicitly addresses emotion preservation in discrete speech representations by integrating emotion-aware optimization at the latent, relational, and alignment levels.

A.2 Emotion-Aware Speech Representation Learning

Emotion-aware speech representation learning has been widely studied in speech emotion recognition and expressive speech modeling. Early approaches (Schuller et al., 2009; Ververidis and Kotropoulos, 2006) relied on handcrafted acoustic features and prosodic descriptors, while later work adopted deep neural networks to learn emotion-discriminative representations directly from waveforms or spectrograms (Yin et al., 2025; Xiao and Das, 2024). More recently, self-supervised speech models such as wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022) have been shown to encode rich affective information and transfer effectively to emotion-related tasks. In parallel, cross-modal contrastive models such as AudioCLIP (Guzhov et al., 2022), CLAP-MST (Elizalde et al., 2023) and CLAP-LAION (Wu et al., 2023b) learn emotion-relevant representations through large-scale audio–text alignment. Building on these representations, several methods (Ma et al., 2024; Wang et al., 2024a; Xiao et al., 2025; Shi et al., 2025c) further adapt or fine-tune pretrained encoders to emphasize emotional attributes, yielding strong performance on standard SER benchmarks.

Beyond recognition, emotion representations have also been incorporated into speech generation systems, particularly emotional and expressive text-to-speech models. Prior work (Wu et al., 2019; Guo et al., 2023; Shi et al., 2025a; Gao et al., 2025) explores disentangling emotion from linguistic content and speaker identity to improve controllability and generalization. However, maintaining stable emotion expression becomes increasingly challenging in complex generation pipelines, especially those involving discrete tokenization and multi-stage synthesis, as emotional cues are often attenuated or distorted during intermediate representation transformations.

Despite these advances, existing emotion-aware representations are typically learned independently of neural speech codecs. In codec-based pipelines (Défossez et al., 2022; Xin et al., 2024b), emotion is commonly treated as an implicit at-

tribute preserved through acoustic reconstruction, without explicit constraints to protect affective structure during discretization. Recent analysis-oriented work (Wu et al., 2024b), such as EmoCodec (Ren et al., 2024), highlights this limitation by systematically evaluating emotion degradation across codecs, but does not modify the codec learning objective itself. This gap motivates approaches that explicitly integrate emotion-aware objectives into discrete speech representation learning, bridging emotion modeling and neural codecs at the representation level.

B Baseline

B.1 Speech Tokenizers and Neural Codecs

EnCodec. EnCodec (Défossez et al., 2022) is an RVQ-based neural audio codec that discretizes speech at a relatively high temporal resolution. It operates at a 75 Hz frame rate and uses two residual codebooks during inference, resulting in a bitrate of approximately 6 kbps. The model is trained with adversarial objectives using multi-scale and multi-period discriminators, and we evaluate it using the official pretrained checkpoint.

SoundStream. SoundStream (Zeghidour et al., 2021) is an end-to-end neural audio codec operating on 24 kHz waveform inputs, using residual vector quantization and adversarial training to get high perceptual quality at low to medium bitrates (3–18 kbps). Quantizer dropout enables bitrate scalability within a single model for real-time streaming.

FunCodec. FunCodec (Du et al., 2024b) is a unified neural audio codec typically operating at 16 kHz, designed to support a wide range of low bitrates (1.5–12 kbps) via a modular RVQ-based framework. It emphasizes flexibility and generalization across compression settings and downstream speech applications.

AudioDec. AudioDec (Wu et al., 2023a) proposes an end-to-end neural audio codec with a modular encoder–quantizer–decoder architecture and a two-stage training strategy that combines reconstruction and adversarial losses. It operates at a frame rate of 50 Hz and supports low-bitrate speech compression in 12.8 kbps, achieving low-latency, high-fidelity reconstruction suitable for streaming scenarios.

AcadmiCodec. AcadmiCodec (Yang et al., 2023) introduces group residual vector quantization (GRVQ) to improve codebook utilization and reconstruction quality under constrained bitrates. It supports frame rates of 50 Hz and 75 Hz, and demonstrates strong performance across a wide bitrate range from 1.5 kbps to 24 kbps, while the accompanying AcademiCodec toolkit provides reproducible training pipelines and pretrained models for neural audio codecs.

MP3, Opus, and AAC. To benchmark emotional trait preservation against established standards, we include widely used conventional codecs as baselines. MP3 (Brandenburg, 1994) provides a strong balance between compression efficiency and audio quality, Opus (Valin et al., 2016) offers broad audio adaptability with low latency, and AAC (Bosi et al., 1997) is commonly adopted for high-fidelity streaming and broadcasting. These codecs serve as reference points for comparison with neural audio codecs.

DAC. DAC (Kumar et al., 2023) extends the VQGAN-style codec framework by projecting latent representations into a low-dimensional space prior to quantization, improving codebook utilization. We evaluate two reproduced variants: one employing three codebooks at a 25 Hz frame rate and another using a single codebook at 75 Hz. Both configurations operate at a 75 Hz token rate and achieve a bitrate of 0.75 kbps, providing a strong low-bitrate acoustic reconstruction baseline.

SpeechTokenizer. SpeechTokenizer (Xin et al., 2024b) enhances discrete speech representations by applying semantic distillation from HuBERT features to the first quantization layer. It operates at a 50 Hz token rate with two codebooks and is designed to improve linguistic modeling for downstream generation. We use the official released checkpoint in all experiments.

Mimi. Mimi (Défossez et al., 2024) follows the hierarchical tokenization design of SpeechTokenizer but replaces the semantic teacher with WavLM representations. It employs eight codebooks, each of size 2,048, and operates at a low frame rate of 12.5 Hz, resulting in a bitrate of approximately 1.1 kbps. This configuration emphasizes compact tokenization with enhanced semantic supervision.

BigCodec. BigCodec (Xin et al., 2024a) explores scaling single-codebook tokenization by increasing model capacity with sequential modules within convolutional architectures. It applies low-dimensional quantization to improve code utilization and operates at an 80 Hz token rate with a codebook size of 8,192, achieving a bitrate of 1.04 kbps.

TAAE. TAAE (Parker et al., 2024) is a transformer-based neural speech codec with nearly 1B parameters that adopts Finite Scalar Quantization (FSQ)

instead of RVQ to enable ultra-low-bitrate compression. Operating at 16 kHz with 25–50 Hz token rates, it achieves 400–700 bps while preserving high perceptual quality.

L1asa. X-Codec 2 (Ye et al., 2025b) adopts a dual-encoder architecture consisting of a semantic encoder based on Wav2Vec2-BERT and an acoustic encoder for low-level features. The outputs of both encoders are concatenated before quantization. The tokenizer operates at a 50 Hz token rate with a large codebook of size 65,536, yielding a bitrate of 0.8 kbps. We use the official pretrained checkpoint.

WavTokenizer. WavTokenizer (Ji et al., 2025) is a single-codebook tokenizer trained on approximately 800K hours of mixed-domain audio. It operates at a 75 Hz token rate with a codebook size of 4,096, resulting in a bitrate of 0.9 kbps. Its large-scale training enables robust performance across diverse acoustic conditions.

B.2 Text-to-Speech Systems

F5-TTS. F5-TTS (Chen et al., 2025) is a flow-matching-based text-to-speech system that directly maps text inputs to acoustic representations without explicit duration modeling. It serves as a strong non-autoregressive baseline for evaluating synthesis quality conditioned on discrete speech tokens.

MaskGCT. MaskGCT (Wang et al., 2024c) is a large-scale masked generative TTS system that removes the need for explicit alignment between text and speech during training. It is trained on the Emilia dataset and relies on masked token prediction to generate speech, enabling flexible zero-shot synthesis.

ARS. ARS (Wang et al., 2024c) is a cascaded autoregressive baseline combining an AR text-to-token model with a non-autoregressive codec-to-waveform decoder. It is also referred to as "AR + SoundStorm" and represents a common two-stage generation paradigm in codec-based TTS systems.

CosyVoice 2. CosyVoice 2 (Du et al., 2024a) is a large-scale zero-shot TTS system built upon an autoregressive language model initialized from Qwen2.5-0.5B-Instruct. It predicts speech tokens extracted by the CosyVoice 2 tokenizer, which operates at a 25 Hz token rate with a bitrate of approximately 0.325 kbps.

FireRedTTS. FireRedTTS (Guo et al., 2024) is an autoregressive TTS system that predicts discrete speech codes extracted by the FireRedTTS tokenizer. The tokenizer employs HuBERT-based semantic features, a ResNet-style encoder, and a single codebook of size 16,384, with decoding performed via flow matching.

SparkTTS. SparkTTS (Wang et al., 2025) is an autoregressive TTS system initialized from Qwen2.5-0.5B-Instruct. It predicts speech tokens produced by the BiCodec tokenizer, enabling an evaluation of BiCodec representations in large-scale generative synthesis.

Llasa. Llasa (Ye et al., 2025b) is a large-scale TTS system built upon an autoregressive model initialized from Llama 3.2-1B. It predicts discrete speech tokens extracted by X-Codec 2, making it particularly relevant for assessing the compatibility of codec representations with large language models.

C Datasets

C.1 Reconstruction Task Datasets

We train the codec on a multi-domain corpus of approximately 2.3K hours of audio, designed to support high-fidelity reconstruction, multilingual robustness, and emotion-aware representation learning. All recordings are resampled to 16 kHz.

LibriSpeech (Panayotov et al., 2015) is used as the primary clean English read-speech source. We include the `train-clean-100` and `train-clean-360` splits, totaling approximately 460 hours. During training, utterances are randomly cropped into 3-second segments. For evaluation, we report reconstruction performance on `test-clean` and `test-other`, representing clean and noisy conditions, respectively.

VCTK (Yamagishi et al., 2019) provides multi-speaker English speech with diverse accents. We use the full corpus (approximately 44 hours), originally recorded at 48 kHz and downsampled to 16 kHz, to improve speaker and phonetic diversity.

AISHELL-3 (Shi et al., 2020) contributes Mandarin speech for cross-lingual robustness. We use the full training set, comprising approximately 85 hours of high-quality recordings at 16 kHz.

AudioSet (Gemmeke et al., 2017) introduces broad acoustic variability. We incorporate a curated 1,000-hour subset covering diverse recording conditions and background environments to improve generalization beyond clean studio speech.

MSP-Podcast (Lotfian and Busso, 2017) is a large-scale spontaneous conversational speech corpus totaling approximately 407 hours of English audio from over 3,600 speakers. Collected from publicly available podcasts, it covers diverse topics, speaking styles, and naturally occurring emotional

expressions, with multi-rater emotion annotations. All audio is provided at 16 kHz, making it well suited for modeling real-world affective speech.

CMU-MOSEI (Bagher Zadeh et al., 2018) is a multimodal sentiment and emotion dataset derived from online videos. We use only the speech modality, comprising approximately 65 hours of English audio with utterance-level emotion annotations. This dataset enriches emotional diversity under unconstrained, in-the-wild recording conditions.

EmoVoiceDB (Yang et al., 2025) is a high-quality English emotional speech dataset designed for fine-grained emotion modeling and evaluation. It contains 40 hours of emotionally expressive speech, comprising over 22,000 utterances annotated with natural language emotion descriptions and covering seven core emotion categories. The dataset includes diverse speaker timbres and expressive styles. All audio is provided at a 16 kHz sampling rate. In our work, EmoVoiceDB is used exclusively as an evaluation benchmark to assess emotional consistency and preservation on reconstructed speech.

C.2 Benchmark for SER Evaluation

EMO-SUPERB (Wu et al., 2024a) aggregates six public SER datasets, including IMPROV, CREMA-D, MSP-Podcast, BIIC-Podcast, IEMOCAP, and NNIME, covering acted, improvised, and real-world emotional speech. In total, the benchmark contains approximately 414 hours of audio from over 2,300 speakers across English and Mandarin. All audio is resampled to 16 kHz and partitioned using speaker-independent splits to ensure robust evaluation. We follow the official EMO-SUPERB data splits and evaluate emotion recognition performance on speech reconstructed by each codec.

C.3 TTS Task Datasets

LibriTTS (Zen et al., 2019) is used to train the downstream text-to-speech token prediction models. LibriTTS is an English read-speech corpus derived from LibriSpeech, providing paired text and speech data with high recording quality. We use the official `train` and `dev` splits, totaling approximately 585 hours of audio. All recordings are resampled to 16 kHz. This dataset serves as the primary source for learning text-conditioned generation over codec tokens.

SECAP (Xu et al., 2024) is a speech emotion captioning dataset containing emotionally expressive utterances annotated with natural language emotion descriptions. The dataset consists of approximately 40 hours of speech from 7 speakers, originally recorded at 24 kHz. We resample the audio to 16 kHz and use it solely for evaluation, focusing on emotional expressiveness and semantic alignment in generated speech rather than caption generation.

D Subjective Reconstruction Evaluation

MUSHRA Evaluation. We conduct a MUSHRA evaluation to assess overall perceptual quality under a reference-based listening protocol (Défossez et al., 2022). Ground-truth recordings are provided as upper anchors, with EnCodec and Llasa included as competitive baselines. A total of 24 listeners participated in the study, and each listener evaluated randomly selected utterances on a 0–100 scale following standard MUSHRA guidelines. As shown in Figure 2 (a), our method achieves a mean MUSHRA score of 90.26, substantially outperforming EnCodec (78.96) and Llasa (87.52), and closely approaching the ground-truth upper bound (91.37). These results indicate that the proposed emotion-guided codec significantly improves perceptual reconstruction quality.

MOS and Emotion-MOS Evaluation. We further assess speech naturalness and affective expressiveness using Mean Opinion Score (MOS) and Emotion-MOS evaluations by fowllowing (Gao et al., 2025). Listeners rated each utterance on a 1–5 scale for overall naturalness and emotional expressiveness, respectively. Each sample was evaluated by 24 independent raters, and scores were averaged across listeners. As illustrated in Figure 2 (b), our method achieves the highest MOS (4.02) and Emotion-MOS (4.21), outperforming both EnCodec (2.92 / 2.67) and Llasa (3.69 / 3.50). The larger improvement in Emotion-MOS highlights the benefit of explicit emotion modeling in discrete representation learning.

AB Preference Evaluation. We additionally conduct AB preference tests to directly compare perceptual and emotional preference between systems. In each trial, listeners were presented with paired samples and asked to indicate overall quality preference and emotional preference. Each comparison was judged by 24 listeners, with randomized order and balanced pairings. As shown in Figure 3 (c), our method is preferred over EnCodec in 78.9% of overall quality judgments and 87.6% of emotional preference judgments, demonstrating a strong lis-

tener preference for the proposed emotion-aware codec.

E Subjective Evaluation for TTS

We conduct subjective listening tests to evaluate perceptual quality and emotional expressiveness of the generated speech using three complementary protocols: Mean Opinion Score (MOS), Emotion Mean Opinion Score (Emotion MOS), and AB preference tests. A total of 24 listeners participated in all evaluations. MOS assesses audio quality and naturalness on a five-point Likert scale ranging from 1 (bad) to 5 (excellent), while Emotion MOS measures the perceived similarity between the emotion expressed in the generated speech and that of the reference audio, rated from 1 (not at all similar) to 5 (extremely similar). For MOS and Emotion MOS evaluations, listeners were asked to rate 30 utterances per system, with emotion categories evenly balanced across samples. The evaluated systems include F5-TTS, CosyVoice 2, and our method. As shown in Figure 3 (a), our method achieves the highest scores on both MOS and Emotion MOS. Specifically, our system attains a MOS of 3.79, outperforming CosyVoice 2 (3.41) and F5-TTS (2.85). A similar trend is observed for Emotion MOS, where our method reaches 4.16, compared to 3.53 for CosyVoice 2 and 2.98 for F5-TTS. The larger margin on Emotion MOS indicates that our approach yields more faithful emotional expression, while also improving overall perceptual quality. We further conduct AB preference tests to directly compare listener preferences between systems. In each trial, listeners were presented with paired samples generated by two systems and asked to select the preferred one based on overall quality and emotional expressiveness, or indicate no preference. Two pairwise comparisons are performed: CosyVoice 2 versus our method and F5-TTS versus our method, using emotion-balanced samples. As shown in Figure 3 (b), listeners prefer our method in 74.7% of comparisons against CosyVoice 2 and 85.5% against F5-TTS, while the proportion of neutral responses remains below 4% in both cases. These results demonstrate a strong and consistent subjective preference for our method, confirming its advantages in both perceptual naturalness and emotional expressiveness during speech synthesis.

F Fine-Grained Ablation Analysis

F.1 Emotion–Semantic Guided Latent

Table 6 reports a fine-grained ablation study on attention–projection strategies for emotion–semantic guided latent modulation, evaluated along emotion consistency, content preservation, and speech naturalness.

The None–Attn variant removes attention and directly injects auxiliary signals by setting $\mathbf{h}_t^{\text{emo}} = \tilde{\mathbf{e}}_t$ and $\mathbf{h}_t^{\text{sem}} = \tilde{\mathbf{s}}_t$, yielding a unified latent $\mathbf{z}_t^{\text{uni}} = \mathbf{z}_t + W_m\tilde{\mathbf{e}}_t + W_m\tilde{\mathbf{s}}_t$. This naive formulation preserves coarse linguistic content with WER 4.58 and WIL 6.60, but exhibits clear degradation in emotional modeling, as reflected by low Emo SIM 0.90 and recall 0.35. In addition, severe spectral distortion is observed, with LSD increasing to 1.05 and MSEP reaching 34.85, indicating that direct feature injection fails to maintain emotion-related structure under quantization.

Single-source supervision partially alleviates this limitation. In the Sem-only variant, emotion modulation is disabled by setting $\mathbf{h}_t^{\text{emo}} = \mathbf{0}$ while retaining $\mathbf{h}_t^{\text{sem}} = \tilde{\mathbf{s}}_t$. This configuration improves spectral stability compared to None–Attn, reducing LSD to 0.83 and MSEP to 25.91, but yields limited gains in emotion consistency, with Emo SIM remaining at 0.89 and recall at 0.34. Conversely, the Emo-only variant sets $\mathbf{h}_t^{\text{emo}} = \tilde{\mathbf{e}}_t$ and $\mathbf{h}_t^{\text{sem}} = \mathbf{0}$, leading to stronger emotion preservation with recall improving to 0.41 and Emo SIM to 0.91, but at the cost of weaker content robustness and higher spectral distortion. These results confirm that emotion and semantic cues provide complementary but insufficient supervision when used independently.

Introducing attention mechanisms further improves representation quality. Self-attention variants replace direct projections with $\mathbf{h}_t^{\text{emo}} = \text{SelfAttn}(\tilde{\mathbf{E}})_t$ and $\mathbf{h}_t^{\text{sem}} = \text{SelfAttn}(\tilde{\mathbf{S}})_t$. Among them, Self–Attn–Before, which applies attention prior to projection, achieves stronger emotion consistency with recall 0.44 and improved content preservation with WER 4.09 and WIL 6.36 compared to Self–Attn–After. It also reduces spectral distortion, attaining LSD 0.80 and MSEP 22.85, suggesting that early interaction in the original feature space is more effective than post-projection refinement. Cross-modal attention yields the most consistent gains by explicitly conditioning acoustic latents on auxiliary signals. The Cross–Attn–After variant applies cross attention after projection and provides moderate improvements in emotion con-

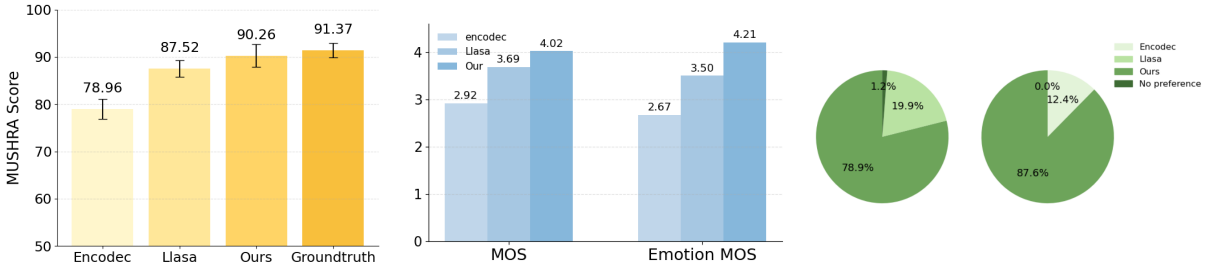

Figure 2: **Reconstruction subjective evaluation results** across three complementary settings. (a) MUSHRA scores comparing Encodec, Llasa, our method, and ground-truth recordings, evaluating overall perceptual quality under a reference-based protocol. (b) MOS and Emotion-MOS results assessing naturalness and affective expressiveness across competing systems. (c) AB-preference results measuring pairwise perceptual preference and emotional preference.

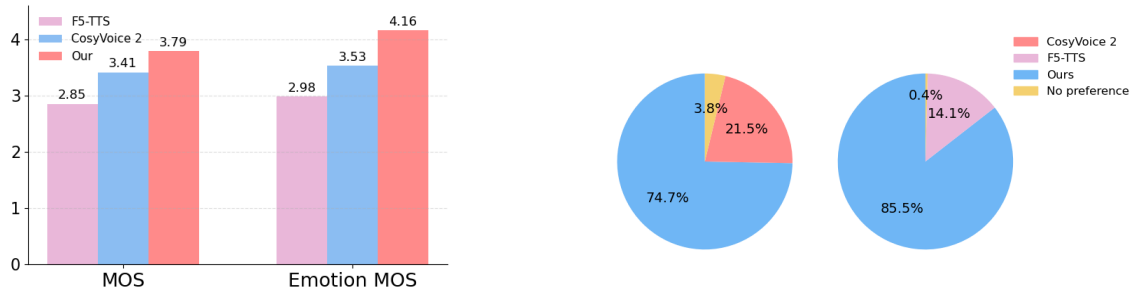

Figure 3: **Text To Speech subjective evaluation results** across two complementary settings. (a) MOS and Emotion-MOS results assessing naturalness and affective expressiveness across competing systems. (b) AB-preference results measuring pairwise perceptual preference and emotional preference.

sistency and perceptual quality. In contrast, the full Cross–Attn–Before configuration applies cross attention prior to projection, enabling richer interaction between acoustic, emotional, and semantic representations in a shared latent space. This design achieves the best overall balance, attaining the highest Emo SIM 0.94 and Pros SIM 0.86, while substantially reducing spectral distortion with LSD 0.78 and MSEP 19.21. Corresponding gains in PESQ 3.04 and UTMOS 3.68 further demonstrate that emotion– semantic guided cross-modal interaction before projection is critical for preserving emotional expressiveness without compromising content fidelity or perceptual naturalness.

F.2 Relation-Preserving Emotional–Semantic Distillation

Table 7 presents a fine-grained ablation study of the proposed relation-preserving emotional–semantic distillation, evaluated from emotion consistency, content preservation, and speech naturalness.

The *None-Distill* variant removes relational supervision entirely, allowing unified latents to be optimized solely by reconstruction objectives. As a result, performance degrades substantially across all dimensions, with notable drops in emotion consistency (Emo SIM 0.92, Recall 0.40) and severe spectral distortion (LSD 1.09, MSEP 40.16). This confirms that residual vector quantization alone is insufficient to preserve relational structure across emotional and semantic spaces.

The *Sem-only* variant introduces relational distillation using only semantic teacher relations by disabling emotional supervision. The corresponding objective is defined as

$$\mathcal{L}_{\text{rela}}^{\text{sem}} = \frac{1}{T'^2} \sum_{t=1}^{T'} \sum_{t'=1}^{T'} \left[\beta \, d\big(r_{t,t'}^{\text{uni}}, r_{t,t'}^{\text{sem}}\big) \right], \quad \alpha = 0. \tag{6}$$

This setting yields clear improvements over *None-Distill* in content preservation (WER 4.34 vs. 4.62, LSD 0.81 vs. 1.09) and speech naturalness (PESQ 2.83, UTMOS 3.52). However, gains in emotion consistency remain limited (recall 0.42), indicating that semantic relations alone cannot fully recover affective structure.

The *Emo-only* variant instead distills relational structure exclusively from the emotion teacher by

Table 6: **Ablation Study of Emotion–Semantic Guided Latent Modulation**. Cross variants incorporate cross-modal attention between emotion and semantic signals, while Self variants apply self-attention. Before applies attention prior to projection into the encoder latent space, whereas After applies attention post-projection. None denotes direct projection without attention. Results are averaged over three random seeds.

Attn–Proj Type	Emotion Consistency			Content Preservation			Speech Naturalness		
	Emo SIM↑	Pros SIM↑	Recall↑	WER↓	WIL↓	LSD↓	MSEP↓	PESQ↑	UTMOS↑
Ours (Cross-Attn-Before)	**0.94**	**0.86**	**0.48**	4.15	6.43	**0.78**	**19.21**	**3.04**	**3.68**
None-Attn	0.90	0.79	0.35	4.58	6.60	1.05	34.85	2.67	3.52
w/ Sem-only	0.89	0.79	0.34	4.67	6.61	0.83	25.91	2.75	3.48
w/ Emo-only	0.91	0.81	0.41	4.72	6.65	0.82	26.40	2.69	3.39
w/ Self-Attn-After	0.90	0.82	0.40	4.67	6.61	0.84	25.10	2.71	3.34
w/ Self-Attn-Before	0.92	0.84	0.44	**4.09**	**6.36**	0.80	22.85	2.89	3.51
w/ Cross-Attn-After	0.91	0.83	0.42	4.17	6.70	0.89	23.90	2.82	3.46

Table 7: **Ablation Study of Relation-Preserving Emotional–Semantic Distillation**.The full RP-Distill is compared with several ablated variants, including None-Distill without relational supervision, Sem-only using semantic relational distillation only, Emo-only using emotion relational distillation only, and Feature Distill applying feature-level distillation without relational constraints. Results are averaged over three random seeds.

Distillation Type	Emotion Consistency			Content Preservation			Speech Naturalness		
	Emo SIM↑	Pros SIM↑	Recall↑	WER↓	WIL↓	LSD↓	MSEP↓	PESQ↑	UTMOS↑
Ours (RP-Distill)	**0.94**	**0.86**	**0.48**	**4.15**	**6.43**	**0.78**	**19.21**	**3.04**	**3.68**
None-Distill	0.92	0.81	0.40	4.62	6.78	1.09	40.16	2.49	3.44
w/ Sem-only	0.92	0.82	0.42	4.34	6.53	0.81	22.58	2.83	3.52
w/ Emo-only	0.93	0.84	0.45	4.60	6.65	0.83	25.14	2.66	3.46
w/ Feature Distill	0.92	0.82	0.42	4.42	6.59	0.82	22.47	2.80	3.49

disabling semantic supervision:

$$\mathcal{L}_{\text{rela}}^{\text{emo}} = \frac{1}{T'^2} \sum_{t=1}^{T'} \sum_{t'=1}^{T'} \Big[\alpha \, d\big(r_{t,t'}^{\text{uni}}, r_{t,t'}^{\text{emo}}\big) \Big], \quad \beta = 0. \tag{7}$$

This configuration substantially improves emotion consistency (Emo SIM 0.93, recall 0.45), but provides weaker benefits for content fidelity and spectral stability, reflected by higher WER (4.60) and LSD (0.83). These results suggest that emotional relations alone are insufficient to stabilize fine-grained acoustic structure under quantization.

The *Feature Distill* variant replaces relational constraints with direct feature matching between unified latents and teacher representations:

$$\mathcal{L}_{\text{feat}} = \frac{1}{T'} \sum_{t=1}^{T'} \left(\left\| \mathbf{z}_t^{\text{uni}} - \mathbf{e}_t \right\|_2^2 + \left\| \mathbf{z}_t^{\text{uni}} - \mathbf{s}_t \right\|_2^2 \right). \tag{8}$$

While feature-level distillation directly aligns unified latents with projected teacher embeddings and improves overall stability compared to *None-Distill*, its performance remains inferior to relation-preserving supervision, particularly in emotion consistency (recall 0.42) and perceptual quality (PESQ 2.80). In contrast, the full relation-preserving emotional–semantic distillation achieves the best overall balance, simultaneously maximizing emotion consistency, reducing spectral distortion, and improving perceptual naturalness. These results demonstrate that preserving pairwise relational geometry across both emotion and semantic spaces is critical for maintaining expressive and content-faithful discrete representations under quantization.

F.3 Emotion-Weighted Semantic Alignment

In Table 8, this ablation study examines how emotion-aware weighting affects semantic alignment under discrete quantization. By progressively removing or simplifying the emotion-dependent weighting mechanism, the three variants isolate the contributions of explicit emotion guidance, semantic-only alignment, and uniform scaling to representation robustness and expressiveness. The *None-Align* variant completely removes the emotion-aware semantic alignment objective, aiming to evaluate codec behavior when discrete representations are trained without explicit semantic alignment. In this setting, the alignment loss $\mathcal{L}_{\text{align}}$ is disabled, and optimization relies solely on re-

Table 8: **Ablation Study of Emotion-Weighted Semantic Alignment.** The full model employs emotion-aware frame weighting during semantic alignment. Sem-only Align removes emotion guidance and relies solely on semantic supervision. Uniform-Scaled Weighted applies uniform weighting without considering emotional variation. Results are averaged over three random seeds.

Distillation Type	Emotion Consistency			Content Preservation			Speech Naturalness		
	Emo SIM↑	Pros SIM↑	Recall↑	WER↓	WIL↓	LSD↓	MSEP↓	PESQ↑	UTMOS↑
Ours (EW-Align)	**0.94**	**0.86**	**0.48**	**4.15**	**6.43**	**0.78**	**19.21**	**3.04**	**3.68**
None-Align	0.93	0.82	0.43	4.51	6.67	1.07	36.59	2.58	3.49
w/ Sem-only Align	0.93	0.83	0.44	4.32	6.49	0.83	24.78	2.84	3.56
w/ Uniform-Scaled Weighted	0.92	0.82	0.42	4.48	6.61	0.97	31.35	2.66	3.50

Table 9: **Ablation of RVQ layer selection.** First Layer supervises only the first RVQ layer $\mathbf{Q}^{(1)}$, whereas Early Layers and All Layers progressively include deeper quantization stages. Across all evaluation dimensions, performance degrades as supervision extends to deeper RVQ layers, indicating that the earliest RVQ layer captures the most compact and informative representation for emotion-aware speech modeling.

RVQ Layer Selection	Emotion Consistency			Content Preservation			Speech Naturalness		
	Emo SIM↑	Pros SIM↑	Recall↑	WER↓	WIL↓	LSD↓	MSEP↓	PESQ↑	UTMOS↑
Ours: First Layer ($\mathbf{Q}^{(1)}$)	**0.94**	**0.86**	**0.48**	**4.15**	**6.43**	**0.78**	**19.21**	**3.04**	**3.68**
w/ Early Layers ($\mathbf{Q}^{(1:4)}$)	0.93	0.84	0.45	4.22	6.42	0.83	22.85	2.89	3.61
w/ All Layers ($\mathbf{Q}^{(1:8)}$)	0.91	0.82	0.41	4.38	6.65	0.95	26.90	2.63	3.52

construction and prior supervisory objectives. As reported in Table 10, removing alignment leads to noticeable degradation in both spectral stability and perceptual quality, with LSD increasing to 1.07 and MSEP rising sharply to 36.59. Although emotion consistency remains moderately high in terms of Emo SIM (0.93), recall drops to 0.43, indicating weaker preservation of emotion-related temporal structure. These results suggest that semantic alignment plays a crucial role in stabilizing discrete representations under quantization, particularly for emotionally expressive speech.

The *Sem-only Align* variant introduces semantic alignment while removing emotion-aware weighting, in order to isolate the effect of semantic supervision alone. The alignment objective reduces to a uniformly weighted formulation:

$$\mathcal{L}_{\text{align}}^{\text{Sem}} = -\frac{1}{T'}\sum_{t=1}^{T'} \log \sigma\Big(\cos(Q_t^{(1)}, C_t^{*})\Big), \quad (9)$$

where all frames contribute equally regardless of emotional variation. Compared to None-Align, Sem-only Align substantially improves content preservation and spectral fidelity, reducing WER to 4.32, WIL to 6.49, and LSD to 0.83, with a large decrease in MSEP from 36.59 to 24.78. However, gains in emotion consistency remain limited, with recall reaching 0.44 and Pros SIM 0.83. This indicates that semantic alignment improves intelligibility and acoustic stability, but lacks the ability to selectively protect emotionally salient regions.

The *Uniform-Scaled Weighted* variant retains emotion-derived weighting but removes temporal adaptivity by replacing frame-wise weights with a global scalar. Specifically,the frame-level emotion weights $\{\gamma_t\}_{t=1}^{T'}$ are averaged into a single constant scaling factor $\bar{\gamma} = \frac{1}{T'}\sum_{t=1}^{T'} \gamma_t$, which is then applied uniformly across all frames, removing temporal emotion variation while preserving the overall weighting magnitude.

$$\mathcal{L}_{\text{align}}^{\text{Uni}} = -\frac{1}{T'}\sum_{t=1}^{T'} \bar{\gamma} \log \sigma\Big(\cos(Q_t^{(1)}, C_t^{*})\Big). \quad (10)$$

This design preserves overall emotion awareness while discarding frame-level discrimination. As shown in Table 10, Uniform-Scaled Weighted improves upon None-Align in content preservation, reducing WER to 4.48 and MSEP to 31.35, and yields slightly better perceptual quality with UTMOS 3.50. However, emotion consistency degrades compared to Sem-only Align, with recall decreasing to 0.42 and LSD remaining relatively high at 0.97. These results demonstrate that emotion-aware weighting alone is insufficient without temporal adaptivity, highlighting the importance of frame-level emotion-sensitive alignment.

F.4 Effect of RVQ layer selection

Table 9 examines the impact of supervising different depths of RVQ layers. In our framework, both the relation-preserving distillation stage (§3.2.2) and the emotion-weighted semantic alignment stage (§3.2.3) apply supervision exclusively to the first RVQ layer $\mathbf{Q}^{(1)} = \{\mathbf{Q}_t^{(1)}\}_{t=1}^{T'} \in \mathbb{R}^{T' \times D}$, which consistently achieves the best performance across emotion consistency, content preservation, and speech naturalness metrics. To study the effect of incorporating deeper quantization stages, we further consider early-layer supervision $\mathbf{Q}^{(1:4)} = \frac{1}{4}\sum_{i=1}^{4} \mathbf{Q}^{(i)} \in \mathbb{R}^{T' \times D}$ and full-layer supervision $\mathbf{Q}^{(1:8)} = \frac{1}{8}\sum_{i=1}^{8} \mathbf{Q}^{(i)} \in \mathbb{R}^{T' \times D}$, where representations from multiple RVQ layers are averaged prior to supervision. Aggregating early RVQ layers leads to moderate degradation, while extending supervision to all RVQ layers further degrades performance across all evaluation dimensions. This monotonic trend suggests that the first RVQ layer captures the most compact and informative structure for emotion-aware speech modeling, whereas deeper RVQ layers predominantly encode residual acoustic details that are less stable and more sensitive to quantization noise. These observations align with prior findings in SpeechTokenizer, which show that RVQ-1 exhibits stronger structural alignment and higher information efficiency than deeper quantization layers (Xin et al., 2024b). Overall, the results indicate that focusing supervision on the earliest RVQ layer is crucial for preserving emotionally and semantically salient structure in discrete speech representations.

F.5 Emotion Encoder Selection

To assess the effect of emotion encoder choice and the robustness of our framework, we evaluate representative encoders that balance emotion modeling and generalization. As shown in Table 12, CLAP-LAION (Wu et al., 2023b) achieves the most balanced overall performance across all metrics, benefiting from its large-scale, language-grounded contrastive training, which promotes both affective generalization and semantic consistency. It attains strong emotion preservation (Emo SIM = 0.94, Pros SIM = 0.86, recall = 0.48) while maintaining content fidelity (WER = 4.15, WIL = 6.43) and speech naturalness (PESQ = 3.04, UTMOS = 3.68). In contrast, CLEP-DG (Shi et al., 2025c) yields a slightly higher Emo SIM (0.95) but shows degraded prosodic consistency and increased distortion, reflecting its reliance on emotion-specific training that limits semantic generalization. CLAP-MST (Elizalde et al., 2023), despite sharing a similar training paradigm, underperforms CLAP-LAION due to its smaller training scale. AudioCLIP (Guzhov et al., 2022) consistently lags behind, particularly in emotion consistency and naturalness, indicating limited suitability for emotion-aware codec supervision. Overall, these results highlight the advantage of large-scale, language-aligned emotion encoders for balanced emotion, content, and prosody preservation.

G Model Details

G.1 Encoder and Decoder Architecture

The proposed codec adopts a standard neural audio codec backbone widely used in recent work on discrete speech representations and neural tokenizers (Défossez et al., 2022; Zeghidour et al., 2021; Xin et al., 2024b). The encoder processes raw waveforms through a hierarchical convolutional architecture to produce temporally downsampled latent representations (Xiao and Das, 2025; Peng and Xiao, 2025). Specifically, the encoder begins with a one-dimensional convolutional layer with 32 channels and a kernel size of 7, followed by four stacked residual convolutional blocks. Each block contains two convolutional layers with kernel size $(3, 1)$ and unit dilation, a residual connection, and a strided convolution for temporal downsampling. The stride factors across the four blocks are set to $(2, 4, 5, 8)$, with kernel sizes twice the corresponding stride. The number of channels is doubled at each downsampling stage to progressively increase representational capacity. To capture long-range temporal dependencies, the convolutional backbone is followed by a two-layer bidirectional LSTM. A final one-dimensional convolution with kernel size 7 projects the hidden states to the target embedding dimension D. ELU (Clevert et al., 2015) activations are used throughout the network, and either layer normalization or weight normalization is applied depending on the layer type. The decoder mirrors the encoder architecture in reverse order. Strided convolutions are replaced with transposed convolutions to restore temporal resolution, and LSTM layers are used to reconstruct long-range structure. The decoder outputs the reconstructed waveform at the original sampling rate.

Table 10: **Ablation of Emotion Encoder Selection.** We evaluate representative emotion encoders designed to balance generalization and emotion preservation. CLAP-LAION demonstrates superior emotion generalization while maintaining strong semantic consistency and prosodic fluency. Results are averaged over three random seeds.

Emo-Encoder Selection	Emotion Consistency			Content Preservation			Speech Naturalness		
	Emo SIM↑	Pros SIM↑	Recall↑	WER↓	WIL↓	LSD↓	MSEP↓	PESQ↑	UTMOS↑
Ours (CLAP-LAION)	0.94	**0.86**	**0.48**	**4.15**	**6.43**	**0.78**	**19.21**	**3.04**	**3.68**
w/ CLEP-DG	**0.95**	0.84	0.47	4.53	6.93	0.90	21.12	2.86	3.57
w/ CLAP-MST	0.92	0.84	0.44	4.26	6.55	0.85	20.42	2.97	3.61
w/ AudioClip	0.90	0.82	0.40	4.82	7.31	0.94	23.03	2.78	3.49

G.2 Residual Vector Quantization

The continuous encoder outputs are discretized using a Residual Vector Quantization (RVQ) module following prior neural codec designs (Défossez et al., 2022; Xin et al., 2024b). RVQ consists of a sequence of K codebooks, where each codebook quantizes the residual error left by the previous ones, enabling progressive refinement of the discrete representation. In our implementation, the encoder latent tensor of shape $[B, D, T]$ is quantized using $K = 8$ residual codebooks, each containing 1024 entries. At each quantization step, the nearest codebook entry is selected and subtracted from the residual, which is then passed to the next codebook. The final discrete representation is obtained by summing the selected embeddings from all codebooks and is fed into the decoder for waveform reconstruction. Codebook entries are updated using exponential moving average updates with a decay factor of 0.99. To prevent codebook collapse, unused entries are periodically replaced by randomly sampled encoder vectors from the current batch. During backpropagation, gradients are propagated through the quantization operation using the straight-through estimator (Bengio et al., 2013).

G.3 Discriminators

To encourage high-fidelity and perceptually realistic reconstructions, adversarial training is employed using multiple discriminators, following common practice in neural codec training (Défossez et al., 2022; Zeghidour et al., 2021; Yang et al., 2023). Three discriminators are used: a Multi-Scale STFT (MS-STFT) discriminator, a Multi-Scale Discriminator (MSD), and a Multi-Period Discriminator (MPD). The MS-STFT discriminator operates on complex-valued short-time Fourier transforms at multiple resolutions, where real and imaginary components are concatenated and processed by a sequence of two-dimensional convolutional layers with increasing temporal dilation. The MSD processes raw waveforms at multiple temporal scales, while the MPD captures periodic structure by reshaping waveforms into two-dimensional representations with different periods following by in (Kong et al., 2020). All discriminators are configured with matched channel dimensions to balance their contributions during adversarial training.

G.4 Training Objective

The training objective combines reconstruction-oriented losses with adversarial supervision to ensure faithful waveform recovery, perceptual naturalness, and stable discrete quantization. Let $\mathbf{x}$ denote the input speech waveform and $\hat{\mathbf{x}}$ its reconstruction produced by the codec.

Quantization Commitment Loss. To stabilize residual vector quantization and prevent codebook collapse, we apply a commitment loss that penalizes the discrepancy between the encoder outputs and their quantized counterparts. Let $\mathbf{z}_j$ denote the residual vector at the j-th quantization stage and $\mathbf{z}_{q_j}$ the selected codebook embedding. The commitment loss is defined as

$$\mathcal{L}_q = \sum_{j=1}^{K} \left\| \mathbf{z}_j - \mathbf{z}_{q_j} \right\|_2^2, \tag{10}$$

where gradients are applied only to the encoder outputs using a straight-through estimator.

Spectral Reconstruction Loss. To encourage spectral fidelity across multiple temporal resolutions, we compute reconstruction loss in the time–frequency domain using mel-spectrograms. Specifically, mel representations $M_i(\cdot)$ are extracted with STFT window sizes 2^i and hop sizes $2^i/4$, where $i \in \{5, \ldots, 11\}$. The spectral loss is defined as

$$\begin{aligned} \mathcal{L}_{\text{mel}} = \sum_i \Big(& \|M_i(\mathbf{x}) - M_i(\hat{\mathbf{x}})\|_1 \\ & + \|M_i(\mathbf{x}) - M_i(\hat{\mathbf{x}})\|_2 \Big) \end{aligned} \tag{11}$$

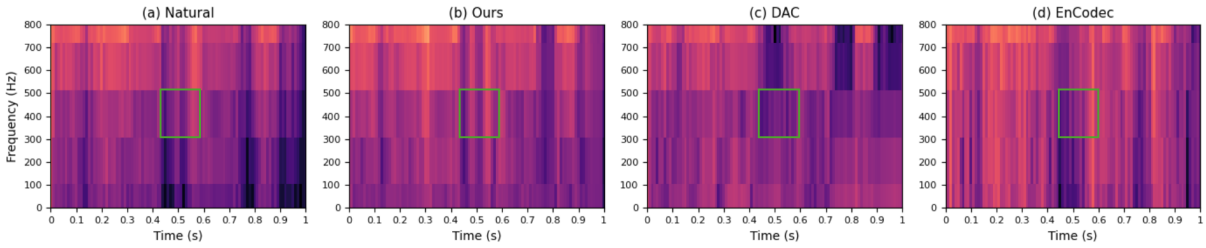

Figure 4: **Qualitative comparison of reconstructed spectrograms across different codecs.** The figure visualizes mel-spectrograms of the same speech segment reconstructed by (a) the natural reference, (b) our method, (c) DAC, and (d) EnCodec. Low-frequency regions associated with prosodic and emotional cues are highlighted for comparison, illustrating differences in temporal continuity and spectral stability across models.

Adversarial Loss. To enhance perceptual realism of reconstructed speech, adversarial supervision is introduced following prior neural codec frameworks (Défossez et al., 2022; Yang et al., 2023). A set of discriminators $\{D_k\}_{k=1}^{K}$ operating at multiple temporal and spectral resolutions is employed, including multi-scale STFT, multi-period, and multi-scale waveform discriminators. The generator is optimized to produce reconstructions that are indistinguishable from real speech. Its adversarial loss is defined using the hinge formulation as

$$\mathcal{L}_{\text{adv}}^{G} = \frac{1}{K}\sum_{k=1}^{K} \max(0,\, 1 - D_k(\hat{\mathbf{x}}))\,, \tag{12}$$

where $\hat{\mathbf{x}}$ denotes the reconstructed waveform. The discriminators are trained to distinguish real speech $\mathbf{x}$ from generated samples $\hat{\mathbf{x}}$. Their objective is given by

$$\begin{aligned}\mathcal{L}_{\text{adv}}^{D} = \frac{1}{K}\sum_{k=1}^{K} \Big[&\max(0,\, 1 - D_k(\mathbf{x})) \\ &+ \max(0,\, 1 + D_k(\hat{\mathbf{x}})) \Big].\end{aligned} \tag{13}$$

This adversarial formulation encourages high-fidelity waveform reconstruction while stabilizing training across multiple discriminative views.

Feature Matching Loss. To stabilize adversarial training and align intermediate representations, we additionally apply a feature matching loss over discriminator activations. Let $D_k^{(l)}(\cdot)$ denote the output of the l-th layer of discriminator k, with L total layers. The feature matching loss is given by

$$\mathcal{L}_{\text{feat}} = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L} \frac{\left\| D_k^{(l)}(\mathbf{x}) - D_k^{(l)}(\hat{\mathbf{x}}) \right\|_1}{\mathbb{E}\left[\left\| D_k^{(l)}(\mathbf{x}) \right\|_1 \right]}, \tag{14}$$

which encourages the generator to match real speech statistics across multiple abstraction levels.

H Qualitative Analysis

Figure 4 provides a qualitative comparison of low-frequency mel spectrograms for natural speech, the proposed method, and two representative neural codecs. The analysis focuses on frequencies below 800 Hz, which are closely linked to prosodic and emotional cues (Singh et al., 2022; Sharan et al., 2024), including fundamental frequency components, lower-order formants, and their temporal dynamics. These regions are particularly sensitive to quantization artifacts and thus offer an informative view of emotional structure preservation. A representative time–frequency region shared across methods (approximately 0.4–0.6 s and 300–500 Hz) is highlighted, where formant transitions and energy modulation are prominent. Natural speech exhibits smooth temporal continuity and coherent spectral evolution in this region. The proposed method closely preserves this behavior, maintaining stable and continuous spectral trajectories across frames. In contrast, DAC and EnCodec display more fragmented patterns, with abrupt temporal variations and reduced local coherence, indicative of quantization-induced disruption in low-frequency spectral structure. Overall, low-frequency spectral regions associated with prosodic and emotional cues show improved temporal continuity under the proposed method, suggesting reduced distortion of affectively salient structure during quantization. These qualitative observations are consistent with the quantitative gains in emotion consistency and perceptual naturalness, providing complementary evidence for the effectiveness of emotion-aware modeling.