X-EMO: DISENTANGLED SPEECH TOKENS FOR CROSS-LINGUAL ZERO-SHOT EXPRESSIVE TEXT-TO-SPEECH

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-lingual zero-shot expressive text-to-speech (TTS) aims to generate speech in a target language while preserving the speaker identity and emotion of a prompt utterance in another language. A key challenge is to transfer these attributes without entangling them with linguistic content and source-language characteristics. We propose X-EMO, a cross-lingual expressive TTS framework built on a disentangled discrete speech tokenizer. During tokenizer training, the decoder receives an additional utterance that shares the speaker and emotion of the input but contains different linguistic content. This makes speaker and emotional information directly available to the decoder, encouraging the discrete representation to primarily encode information required for linguistic reconstruction. The resulting tokens are modeled by an autoregressive language model, while a flow-matching acoustic model synthesizes speech conditioned on speaker and emotion embeddings from the prompt utterance. With approximately 65 hours of expressive speech added to bilingual read-speech data, X-EMO achieves competitive intelligibility, speaker similarity, naturalness, and emotion preservation across multiple languages, while maintaining stable cross-lingual performance across diverse prompt languages.

1 INTRODUCTION

Cross-lingual expressive text-to-speech (TTS) aims to synthesize natural speech in a target language while preserving the speaker identity and emotional characteristics of a prompt utterance from another language (Gudmalwar et al., 2024; Zhu et al., 2024). This capability is important for multilingual speech applications, such as speech-to-speech translation, where preserving emotion information is essential for maintaining the richness of the original speech beyond its linguistic content (Heffernan et al., 2024). However, cross-lingual expressive TTS remains relatively underexplored, with existing approaches largely focusing on a limited set of resource-rich languages (Gong et al., 2025; Liu et al., 2026; Zheng et al., 2026; Zuo et al., 2025; Chen et al., 2026). Extending expressive speech synthesis to a broader range of languages therefore remains challenging, particularly when expressive speech data are scarce (Hosseini-Kivanani & Dowerah, 2026; Cong et al., 2023).

Disentangling speech representations provides a promising way to reduce the reliance on expressive speech data. Early work such as AutoVC (Qian et al., 2019) showed that an information bottleneck can enable zero-shot voice style transfer without explicit disentanglement supervision. Subsequent approaches, including Vevo (Zhang et al., 2025) and NaturalSpeech 3 (Ju et al., 2024), further explored self-supervised and factorized representations for expressive speech synthesis. However, these approaches typically rely on predefined speech factors such as linguistic content, timbre, and prosody. For cross-lingual expressive TTS, accent, emotion, and speaker identity may be intertwined across these factors, making such predefined factorization insufficient for isolating these attributes. Moreover, many disentanglement methods rely on information bottlenecks (Cong et al., 2023; Zhang et al., 2023a), such as limiting the bottleneck dimension or codebook size. These methods offer only coarse control over the information retained in the representation, and the attributes to be removed cannot be explicitly specified or controlled (Qian et al., 2019; Bous & Roebel, 2023).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address this challenge, we propose X-EMO, a language-model and flow-matching-based TTS system built on a disentangled speech tokenizer. We draw on the information preference principle of Variational Lossy Autoencoders (Chen et al., 2016). As shown in Fig. 1, during training of the disentangled speech tokenizer, we provide a conditioning utterance to the decoder as well as the input utterance fed to the encoder. The conditioning utterance has the same speaker and emotion as the input but different linguistic content. The decoder can therefore obtain speaker and emotion information directly from the conditioning utterance. Following the information preference principle, the discrete latent representation is encouraged to retain only information unavailable from the condition, while filtering out information already provided to the decoder. This encourages the tokens to discard speaker and emotion while retaining linguistic information. We then use a language model to model these tokens and a flow-matching acoustic model to generate speech from the predicted tokens conditioned on speaker and emotional embeddings.

This design decouples linguistic modeling from speaker and emotion modeling. At inference time, we use a native-speaker anchor utterance in the target language as the language model prompt. Conditioned on this anchor, the language model predicts linguistic tokens for the target text. Only the target-text tokens are passed to the flow-matching model. The flow-matching model combines these tokens with speaker and emotion embeddings extracted from a cross-lingual prompt utterance. Since the flow-matching model does not require a transcription of the prompt speech, our system avoids the additional transcription burden for cross-lingual prompts.

Our approach makes three contributions. First, we introduce a conditional discrete speech tokenizer that uses an attribute-matched utterance with different linguistic content during reconstruction. By making speaker and emotion information available to the decoder, the information cost discourages the discrete representation from redundantly encoding these attributes, while preserving information necessary for reconstruction. Second, we decouple linguistic realization from speaker and emotion control in a two-stage TTS architecture: the language model predicts the disentangled tokens with a target-language anchor, while a flow-matching acoustic model injects speaker and emotion from a separate prompt utterance. Third, we demonstrate the effectiveness of this design on four languages under limited expressive-data supervision, with controlled ablations on the information cost and tokenizer training data.

2 Related Work

Cross-lingual emotional TTS. Early cross-lingual emotional TTS systems disentangle emotion from speaker identity and linguistic content within conventional architectures based on continuous representations, e.g., through cross-speaker and cross-lingual emotion transfer (Li et al., 2023; Zhu et al., 2024), controllable voice identity and emotional style (Gudmalwar et al., 2024), stage-wise modeling of emotion and speaker (Gong et al., 2025), or retrieval-augmented intensity control (Zuo et al., 2025). More recently, language model-based TTS systems have achieved strong multilingual and expressive generation (Anastassiou et al., 2024; Du et al., 2025; Hu et al., 2026). IndexTTS-2.5 (Li et al., 2026) enables independent control of speaker identity and emotion, and Confucius4-TTS (Wang et al., 2026) supports cross-lingual voice and emotion transfer across 14 languages. These capabilities largely stem from large-scale training data, which is less accessible for low-resource languages, where expressive speech is particularly scarce.

Disentangled speech representations. Discrete speech tokens used in language model-based TTS (Wang et al., 2023a; Zhang et al., 2023b) tend to jointly encode linguistic and paralinguistic attributes (Zhang et al., 2024; Jia et al., 2025). One line of work limits the retained information through bottlenecks on dimension, codebook size, or temporal resolution (Qian et al., 2019; 2020; Bous & Roebel, 2023; Zhang et al., 2023a); such bottlenecks control *how much* information is kept, but not *which* attributes are removed. Another line factorizes speech into predefined components such as content, timbre, and prosody (Ju et al., 2024; Zhang et al., 2025), or distills semantic information from self-supervised or ASR models into the tokens (Zhang et al., 2024; Ye et al., 2025; Du et al., 2025). These approaches assume that the attributes to be removed align with fixed factors, whereas in cross-lingual expressive TTS, accent, emotion, and speaker identity may span several of them.

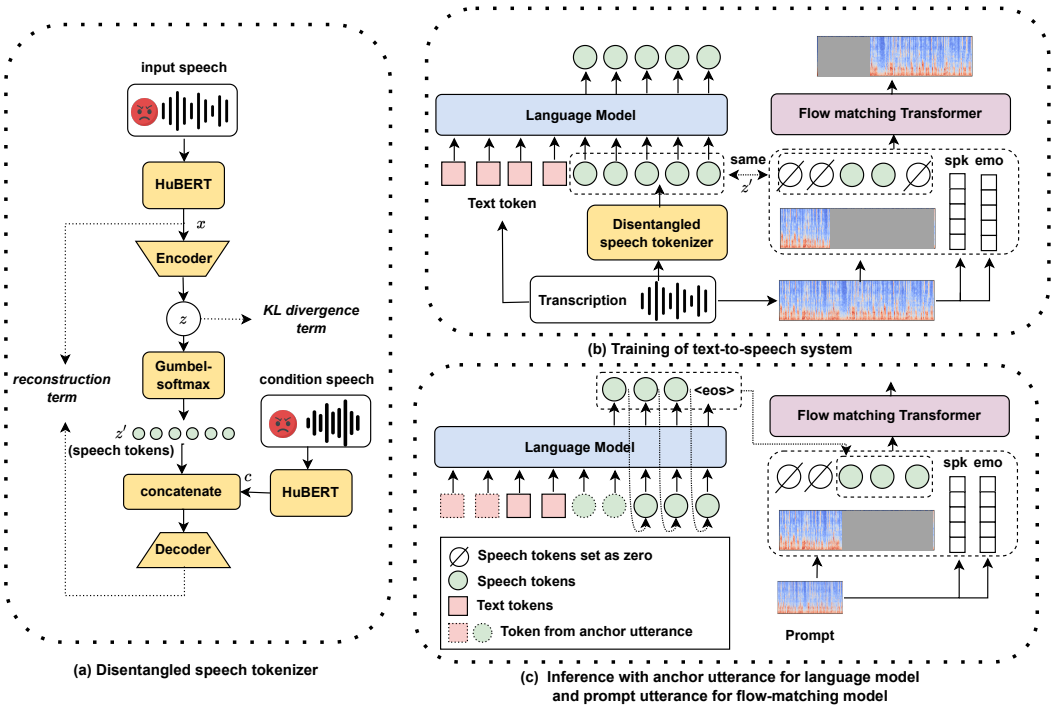


Figure 1: Overview of X-EMO. (a) The disentangled speech tokenizer learns speech tokens that disentangle linguistic content from speaker and emotion attributes. (b) During training, the disentangled speech tokens are used to train the language model and the flow-matching acoustic model. (c) During inference, a native-speaker anchor utterance is provided to the language model for in-context learning, while a prompt utterance provides the speaker and emotion attributes to be preserved by the flow-matching model.

3 Methodology

3.1 Disentangled Speech Tokenizer

Design Principle. Our tokenizer is designed based on the information preference principle of Variational Lossy Autoencoders (Chen et al., 2016; Qian et al., 2020). The principle states that a Variational Autoencoder (VAE) tends to encode in its latent representation only information that cannot be obtained from the decoder input, while discarding information already available from it. Formally, this principle is realized by maximizing an objective that balances reconstruction against the information cost:

$$\mathcal{L}_{\text{ELBO}} = \underbrace{\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z,c)]}_{\text{reconstruction term}} - \beta \underbrace{D_{\text{KL}}(q_\phi(z|x)\|p(z))}_{\text{KL divergence term}} \tag{1}$$

where x is the input signal, z is the latent representation, c is the conditioning signal, $q_\phi(z|x)$ is the encoder, and $p_\theta(x|z,c)$ is the decoder. Here $p(z)$ is the prior distribution over the latent representation. The reconstruction term encourages faithful input reconstruction, while the KL divergence term constrains the information capacity of the latent. We refer to the KL divergence term as the information cost because encoding information already available from c incurs additional KL cost without improving reconstruction.

In our setting, x is the input utterance, while c is a separate utterance sharing its speaker and emotion but with different linguistic content. We therefore expect the speaker and emotion information to be provided by c and filtered out from the latent representation.

Information cost. VQ-VAE (Van Den Oord et al., 2017) and its variants (Défossez et al., 2022) have been widely adopted in TTS to obtain discrete speech tokens (Wang et al., 2023a; Zhang

et al., 2025; Wang et al., 2025). However, a categorical posterior over K codes combined with a uniform prior yields a constant KL value of $\log K$, which cannot serve as a tunable information cost (Van Den Oord et al., 2017). We instead use a relaxed categorical posterior with Gumbel-Softmax quantization (Baevski et al., 2020; Jang et al., 2016). Specifically, we first project z to the codebook dimension:

$$z_e = Wz + b, \tag{2}$$

where W and b are learnable projection parameters. Let $E \in \mathbb{R}^{K \times d}$ denote the codebook, whose k-th entry is E_k. We compute logits $\ell \in \mathbb{R}^K$ from the cosine similarity between z_e and each codebook entry:

$$\ell_k = \frac{z_e^\top E_k}{\|z_e\| \, \|E_k\|}. \tag{3}$$

The categorical posterior is defined as

$$q_\phi(k \mid x) = \mathrm{softmax}(\ell), \tag{4}$$

which is used to compute the KL term against the prior $p(z)$. During training, we obtain the quantized latent z' via a differentiable Gumbel-Softmax sample over the codebook, allowing gradients to flow through the discrete bottleneck. The tokenizer is trained with the following objective:

$$\mathcal{L} = \|x - \hat{x}_\theta(z', c)\|_1 + \beta D_{\mathrm{KL}} \left(q_\phi(k \mid x) \| p(k) \right), \tag{5}$$

where $\hat{x}_\theta(z', c)$ denotes the reconstruction produced by the decoder, β controls the strength of the information cost, and $p(k)$ is a uniform prior over the K codebook entries.

Tokenizer Architecture and Conditioning. The tokenizer comprises an encoder, a discrete bottleneck, and a decoder. Following Zhang et al. (2025), we use the 18th-layer representations of HuBERT (Hsu et al., 2021) as the input representations for both x and c. Accordingly, x and c denote these HuBERT feature sequences, and the decoder reconstructs the HuBERT features of x. During training, we construct (x, c) pairs such that c shares the speaker identity and emotion of x while differing in linguistic content. Importantly, c is used only during tokenizer training. At inference time, speech tokens are extracted from the encoder alone.

3.2 Text-to-token Language Model

The tokenizer is designed to encourage the discrete representation to primarily encode linguistic content while reducing speaker and emotion-related information. We model these tokens with a decoder-only autoregressive language model. As shown in Fig. 1 (b), during training, the language model generates a sequence of speech tokens, where the ground truth speech tokens are also used to condition the flow-matching decoder for mel-spectrogram generation. We follow the standard text-to-token formulation used in related works (Ye et al., 2025; Wang et al., 2023a; Zhang et al., 2023b), with next-token cross-entropy applied to the speech-token sequence.

Target-language anchor. we provide a target-language utterance as an *anchor* prompt for cross-lingual synthesis as shown in Fig. 1 (c). The anchor provides target-language realization cues, such as pronunciation and prosody, while speaker identity and emotion are supplied separately to the flow-matching decoder through the prompt utterance. the language model generates speech tokens autoregressively conditioned on the text tokens and the anchor prompt. Then tokens generated for the target text are passed to the flow-matching model. This design reduces the transfer of source-language accents to the synthesized speech. In practice, we use a single fixed native-speaker anchor utterance per target language; when the prompt and target languages are the same, the prompt utterance itself serves as the anchor, following the standard zero-shot TTS setting.

3.3 Flow-matching acoustic model

Given the speech tokens generated by the autoregressive model, we use a flow-matching transformer to generate the corresponding mel-spectrogram. The mel-spectrogram is extracted at the same 50Hz frame rate as the speech tokens, allowing the two representations to be directly aligned along the temporal dimension. In addition to the speech tokens, the model is conditioned on language identity, speaker identity, emotion, and an acoustic prefix. The speech tokens provide frame-level content information, while language identity is provided at the frame level following the dual-level

language conditioning strategy of X-Voice (Xu et al., 2026). Speaker identity and emotion are injected as utterance-level conditions. Speaker and emotion embeddings are extracted from the prompt utterance using a pretrained CAM++ speaker verification model (Wang et al., 2023b) and Emotion2Vec (Ma et al., 2024), respectively. During training, the acoustic prefix is taken from a clean segment of the target mel-spectrogram, while during inference it is extracted from the prompt audio. The speech tokens corresponding to the acoustic prefix are set to zero during both training and inference. Since speech tokens and mel frames are temporally aligned, this masking can be applied directly without additional alignment.

We use optimal-transport conditional flow matching. Given a target mel-spectrogram x_1 and Gaussian noise $x_0 \sim \mathcal{N}(0, \mathbf{I})$, we construct an intermediate state x_t along a linear probability path:

$$x_t = (1 - (1 - \sigma)t)x_0 + tx_1, \tag{6}$$

where $\sigma = 10^{-5}$ and $t \in [0, 1]$. The flow-matching transformer parameterized by θ takes the intermediate state x_t, the flow step t, and the conditioning signals $\mathbf{y}$ as input, and predicts the velocity field u:

$$\begin{aligned} u &= x_1 - (1 - \sigma)x_0, \\ \mathcal{L}_{\mathrm{FM}} &= \mathbb{E}_t \left[\| v_\theta(x_t, t, \mathbf{y}) - u \|_2^2 \right], \end{aligned} \tag{7}$$

where the loss is computed only on non-prefix, non-padded frames.

4 EXPERIMENTAL SETUP

4.1 TRAINING DATA

The tokenizer is trained on approximately 2k hours of bilingual read speech, comprising 960 hours of English speech from LibriSpeech (Panayotov et al., 2015) and 950 hours of Mandarin speech collected from publicly available speech corpora (Yang et al., 2022; Bu et al., 2017; Ardila et al., 2020). We additionally incorporate approximately 65 hours of emotional speech from MSP-Podcast (Busso et al., 2025), ESD (Zhou et al., 2021), CREMA-D (Cao et al., 2014), and IEMOCAP (Busso et al., 2008) to expose the tokenizer to diverse emotional expressions.

The language model and flow-matching model are trained in two regimes. Ablation study is conducted in a resource-constrained regime of approximately 5k hours of English and Mandarin speech, chosen to emulate a low-resource language and to study the effect of key design choices. The configuration selected in this regime is then trained at scale on approximately 100k hours of multilingual speech spanning English, Mandarin, Malay and Indonesian.

4.2 EVALUATION DATA

We evaluate our system from three perspectives: multilingual TTS, cross-lingual TTS, and emotion preservation. For multilingual TTS, we consider English, Mandarin Chinese, Indonesian, and Malay, using the official Seed-TTS (Anastassiou et al., 2024) test sets for English and Chinese, the Indonesian subset of the X-Voice test set (Xu et al., 2026), and a Malay test set constructed from Common Voice Spontaneous Speech 4.0 (Mozilla Data Collective, 2026). For cross-lingual TTS, we construct test pairs from these multilingual test sets by replacing each prompt with a randomly selected prompt from another language while keeping the target text unchanged. For emotion preservation, we use emotion-labeled speech from the CosyVoice 3 Evaluation (Gao et al., 2025; Du et al., 2025) for English and Chinese, INESCO (Anggrayni et al., 2026) for Indonesian, and the same Malay data used for multilingual TTS. We construct evaluation pairs by combining emotion-containing prompts with target texts across the four languages, covering both monolingual and cross-lingual emotion cloning. Since no Malay speech data with emotion labels is available, Malay is evaluated only as a target language, using emotion-containing prompts from the other three languages.

4.3 METRICS

We follow the evaluation protocol of Seed-TTS for intelligibility and speaker preservation. We report word error rate (WER) for English, Indonesian, and Malay, and character error rate (CER) for

Table 1: Ablation study on the β. Each generation direction denotes the language of the prompt utterance followed by the language of the generated target text (e.g., en→zh indicates an English prompt utterance and Chinese target text). The final two columns report the average WER/CER and Speaker Similarity across the four generation directions. Bold indicates the best result.

	en→en		**zh→zh**		**en→zh**		**zh→en**		**Avg.**	
β	WER ↓	SIM ↑	CER	SIM	CER	SIM	WER	SIM	WER/CER	SIM
0	7.95	0.564	7.29	0.691	8.70	0.392	12.17	0.471	9.03	0.530
0.01	7.65	0.566	5.78	0.698	7.43	0.398	7.54	**0.483**	7.10	0.536
0.05	**7.03**	0.567	**4.92**	**0.705**	**6.27**	0.405	**6.40**	0.479	**6.16**	**0.539**
0.07	7.61	**0.569**	5.53	0.702	6.76	**0.406**	7.31	0.475	6.80	0.538
0.10	7.61	0.562	5.56	0.704	7.44	**0.406**	7.68	0.477	7.07	0.537

Chinese. Whisper-large-v3 (Radford et al., 2022) is used for non-Chinese languages and Paraformer-zh (Gao et al., 2023) is used for Chinese. Speaker similarity (SIM) is measured using the fine-tuned WavLM-large speaker verification model (Chen et al., 2022b;a). Additionally, we use UTMOSv2 (Baba et al., 2024) to assess the naturalness of the generated speech. For emotion preservation, we use Emotion2Vec (Ma et al., 2024) to perform emotion classification. We report emotion accuracy (Acc) and weighted F1 (W-F1), together with the F1 scores for the three emotions of happy, angry, and sad following the evaluation protocol in Li et al. (2026).

5 Results

5.1 Ablation study: Effect of the strength of the information cost

We investigate the effect of the strength of the information cost β in Eq. 5. As shown in Table 1, introducing a moderate information cost substantially improves linguistic accuracy. The average WER/CER decreases from 9.03 with $\beta = 0$ to 7.10 with $\beta = 0.01$, and further to 6.16 with $\beta = 0.05$. In contrast, the change in speaker similarity is relatively small, with the average SIM varying from 0.530 to 0.539 across the tested values.

This difference is consistent with the architecture of our system. The information cost is imposed on the discrete representation used by the language model, and therefore primarily controls the amount of information that this module needs to encode. In contrast, speaker identity is explicitly provided to the subsequent flow-matching acoustic model through speaker conditioning, and β does not directly affect this speaker representation. Consequently, changing β has a much stronger effect on linguistic accuracy than on speaker similarity.

Interestingly, increasing β beyond 0.05 leads to degraded linguistic accuracy, with the average WER/CER increasing to 6.80 and 7.07 for $\beta = 0.07$ and 0.10, respectively. This suggests that an appropriate information cost is beneficial because it reduces the amount of non-essential information that the language model needs to model, while an overly strong constraint can suppress information necessary for linguistic content. We therefore use $\beta = 0.05$ in the remaining experiments.

5.2 Ablation study: Effect of Expressive Data in Tokenizer Training

We investigate the effect of including expressive speech during tokenizer training. Without expressive speech, the conditioning utterance primarily provides speaker information, encouraging the tokenizer to remove speaker identity from the discrete representation. By introducing expressive speech, the conditioning utterance also provides expressive information, allowing the information cost to encourage its removal from the discrete representation as well.

As shown in Table 2, training with the expressive corpus consistently improves both linguistic accuracy and speaker similarity across all four generation directions. On average, WER/CER decreases from 8.00 to 6.16, while speaker similarity increases from 0.529 to 0.539. This suggests that exposure to expressive speech during training helps the learned representation better accommodate acoustic and prosodic variation while preserving the information required for linguistic reconstruction.

Table 2: Ablation on expressive speech in tokenizer training. w/o Expressive and w/ Expressive denote training the tokenizer without and with the expressive corpus, respectively. Bold indicates the better result within each pair.

	en→en		zh→zh		en→zh		zh→en		Avg.	
Training	**WER**	**SIM**	**CER**	**SIM**	**CER**	**SIM**	**WER**	**SIM**	**WER/CER**	**SIM**
w/o Expressive	7.38	0.559	6.19	0.695	7.85	0.396	10.58	0.465	8.00	0.529
w/ Expressive	**7.03**	**0.567**	**4.92**	**0.705**	**6.27**	**0.405**	**6.40**	**0.479**	**6.16**	**0.539**

Table 3: Ablation on expressive speech in tokenizer training for emotion preservation. w/o Exprs. and w/ Exprs. denote training without and with the expressive corpus, respectively. Higher is better for all metrics. Bold indicates the better result within each pair.

	en→en		zh→zh		en→zh		zh→en	
Metric	**w/o Exprs.**	**w/ Exprs.**	**w/o Exprs.**	**w/ Exprs.**	**w/o Exprs.**	**w/ Exprs.**	**w/o Exprs.**	**w/ Exprs.**
Acc	0.540	**0.560**	0.627	**0.647**	0.380	**0.387**	**0.593**	0.520
W-F1	0.642	**0.683**	0.718	**0.732**	0.413	**0.455**	**0.643**	0.571
$F1_{\text{ang}}$	**0.774**	0.732	0.736	**0.764**	0.707	**0.733**	**0.756**	0.651
$F1_{\text{hap}}$	0.721	**0.725**	0.796	**0.820**	0.465	**0.554**	**0.696**	0.649
$F1_{\text{sad}}$	0.431	**0.592**	**0.622**	0.611	0.068	**0.077**	**0.478**	0.413

We further evaluate emotion preservation ability in Table 3. The effect is different across language directions. Expressive training improves most metrics in three of the four directions, whereas zh→en shows a decrease in all five metrics. Interestingly, zh→en also has the largest WER improvement in Table 2. This shows that better linguistic accuracy does not necessarily lead to better emotion preservation. One possible explanation is that speaker identity and emotion have different information structures. Speaker identity is relatively stable within an utterance. It can therefore be represented by a global speaker embedding. The acoustic model can use this embedding to control speaker identity. However, emotion is only partially captured by global utterance-level attributes, while its finer-grained realization is manifested at the frame level. This may explain why the improvement in linguistic accuracy does not always translate into better emotion preservation. The results may also suggest that our tokenizer can filter out only part of the global emotion information. We leave further investigation of this disentanglement to future work.

5.3 Cross-Lingual Emotion Preservation

We evaluate whether the emotion conveyed by the prompt utterance can be preserved when generating speech in a different target language. For completeness, we use the same baselines across directions whenever possible; however, the baseline set may vary across directions when a system does not support the corresponding language pair. The baselines shown here include Confucius4-TTS (Wang et al., 2026), X-Voice (Xu et al., 2026), OmniVoice (Zhu et al., 2026), CosyVoice3 (Du et al., 2025), VoxCPM2 (Zhou et al., 2026). As shown in Table 4, our model achieves the highest emotion accuracy in most cross-lingual directions. The advantage is reflected not only in overall accuracy and weighted F1, but also in the per-emotion F1 scores, indicating that the model does not achieve high overall performance by favoring a particular emotion. This observation is consistent with our hypothesis that the discrete speech representation can primarily encode linguistic information while emotional attributes are provided through the prompt utterance, supporting the effectiveness of our proposed disentanglement strategy.

5.4 Comparison with Multilingual TTS Baselines

Table 5 compares our model with recent multilingual TTS baselines across English, Chinese, Malay, and Indonesian under different language combinations. In the main paper, we include only baselines that support all four languages, enabling a consistent comparison across the same set of language directions. The complete results across more baselines are shown in the Appendix. Baselines shown here are the same as those used in the previous section, except for MOSS-TTS (Gong et al., 2026).

Table 4: Emotion preservation results for cross-lingual generation. $F1_{ang}$, $F1_{hap}$ and $F1_{sad}$ stand for F1 scores for the emotions of angry, happy and sad respectively. The direction denotes the language of the prompt utterance followed by the language of the target text, i.e., *prompt language → target language*. Bold and underline indicate the best and second-best results, respectively.

Direction	System	Acc	W-F1	$\mathbf{F1}_{\mathrm{ang}}$	$\mathbf{F1}_{\mathrm{hap}}$	$\mathbf{F1}_{\mathrm{sad}}$
en → id	Confucius4-TTS	0.453	0.517	0.766	0.361	0.424
	X-Voice	0.340	0.434	0.625	0.465	0.211
	OmniVoice	0.487	0.569	0.674	0.505	0.526
	Ours	**0.647**	**0.718**	**0.800**	**0.713**	**0.640**
zh → id	Confucius4-TTS	0.520	0.592	0.722	0.642	0.413
	X-Voice	0.507	0.555	0.651	0.708	0.305
	OmniVoice	0.493	0.573	0.553	0.667	0.500
	Ours	**0.707**	**0.743**	**0.808**	**0.893**	**0.529**
id → en	Confucius4-TTS	0.653	0.714	0.849	0.651	0.641
	X-Voice	0.513	0.606	0.710	0.629	0.478
	OmniVoice	**0.747**	**0.805**	**0.891**	**0.659**	**0.864**
	Ours	0.653	0.731	0.863	0.638	0.692
zh → en	CosyVoice3	0.360	0.434	0.295	0.596	0.412
	X-Voice	0.473	0.550	0.682	0.608	0.361
	Confucius4-TTS	0.613	0.685	**0.740**	0.723	0.592
	Ours	**0.647**	**0.681**	0.682	**0.740**	**0.622**
id → zh	Confucius4-TTS	**0.507**	0.598	**0.903**	0.200	0.691
	X-Voice	0.313	0.440	0.667	0.190	0.462
	OmniVoice	0.460	0.577	0.809	0.237	0.684
	Ours	0.500	**0.630**	0.791	**0.382**	**0.718**
en → zh	CosyVoice3	0.193	0.302	0.438	0.358	0.109
	X-Voice	0.180	0.286	0.387	0.364	0.107
	Confucius4-TTS	0.387	0.484	**0.822**	0.400	0.230
	Ours	**0.400**	**0.529**	0.675	**0.667**	**0.246**
en → ms	Confucius4-TTS	0.380	0.480	0.674	0.430	0.336
	OmniVoice	0.352	0.472	0.627	0.439	0.349
	VoxCPM2	0.458	0.559	0.675	0.589	0.413
	Ours	**0.563**	**0.665**	**0.707**	**0.675**	**0.613**
zh → ms	Confucius4-TTS	0.588	0.644	**0.633**	0.745	0.553
	OmniVoice	0.552	0.630	0.552	0.730	0.609
	VoxCPM2	0.579	0.644	0.586	0.725	0.620
	Ours	**0.657**	**0.729**	0.632	**0.832**	**0.724**
id → ms	Confucius4-TTS	0.433	0.525	0.698	0.321	0.556
	OmniVoice	0.488	**0.603**	0.671	**0.407**	**0.733**
	VoxCPM2	**0.501**	0.562	**0.782**	0.259	0.645
	Ours	0.430	0.533	0.540	0.359	0.698

Our model does not match the strongest baselines in same-language intelligibility or speaker similarity. However, these systems are trained on 5–20 times more data. The main advantage of our approach lies in the stability of intelligibility across prompt languages: for Indonesian targets, our WERs with English, Chinese, and Malay prompts are 3.12, 3.09, and 3.10, whereas OmniVoice, MOSS-TTS, and VoxCPM2 are much higher. Confucius4-TTS shows similar stability, but it is trained on approximately five times more data. We note that our same-language results on Malay and Indonesian (7.46 and 5.83 WER) are worse than our cross-lingual results. In the same-language setting, the prompt itself serves as the language-model anchor, whereas cross-lingual generation uses a fixed native-speaker anchor. Since the Malay test prompts are drawn from spontaneous speech,

Table 5: Multilingual TTS results on English, Chinese, Malay, and Indonesian. Each column corresponds to a prompt language, and each gray row indicates the target language; thus, each cell represents a *prompt→target* generation direction. We report WER/CER, speaker similarity (SIM), and UTMOSv2 (MOS). Bold and underline indicate the best and second-best results, respectively.

System (Params, Hours)	Prompt Language											
	English			Chinese			Malay			Indonesian		
	WER/CER↓	SIM↑	MOS↑	WER/CER	SIM	MOS	WER/CER	SIM	MOS	WER/CER	SIM	MOS
Target Language: English												
Confucius4-TTS (0.5B, 500K)	**1.44**	0.70	2.86	**1.80**	0.58	2.85	2.45	0.68	2.37	2.47	0.60	2.66
OmniVoice (0.8B, 581K)	1.66	0.74	**3.02**	2.41	**0.67**	2.87	2.49	0.73	2.59	2.06	0.60	**2.98**
MOSS-TTS (8B, 1M)	2.38	0.69	3.00	3.06	0.59	**2.96**	5.56	0.69	**2.62**	8.12	0.53	2.87
VoxCPM2 (2B, 2M+)	1.61	**0.75**	2.72	2.39	0.66	2.65	4.13	**0.75**	2.29	3.04	**0.68**	2.53
Ours (0.7B, 100K)	2.21	0.68	2.81	2.29	0.58	2.86	**2.22**	0.52	2.58	**2.02**	0.43	2.73
Target Language: Chinese												
Confucius4-TTS	**1.23**	0.51	2.63	1.01	0.76	2.70	1.83	0.68	2.31	1.80	0.58	2.60
OmniVoice	1.46	0.52	**2.79**	**0.87**	0.78	2.65	7.02	0.74	2.50	**1.52**	0.56	**2.84**
MOSS-TTS	2.41	0.47	2.76	1.30	0.76	**2.81**	8.90	0.70	**2.53**	6.65	0.52	2.79
VoxCPM2	2.41	**0.54**	2.49	1.24	**0.79**	2.45	7.94	**0.76**	2.23	7.98	**0.65**	2.57
Ours	2.51	0.50	2.61	2.56	0.76	2.64	**1.80**	0.68	2.34	1.79	0.54	2.45
Target Language: Malay												
Confucius4-TTS	**3.17**	0.59	2.67	**2.26**	0.61	2.71	**2.24**	0.78	2.41	**2.53**	0.70	2.45
OmniVoice	6.62	0.63	**2.89**	7.65	**0.71**	2.77	2.79	**0.82**	2.58	2.64	0.72	**2.69**
MOSS-TTS	4.48	0.50	2.76	5.12	0.59	**2.81**	5.37	0.80	**2.61**	6.02	0.65	2.65
VoxCPM2	6.32	**0.66**	2.53	5.02	0.70	2.51	3.21	0.82	2.30	2.72	**0.75**	2.33
Ours	3.21	0.46	2.63	3.30	0.63	2.73	7.46	0.79	2.30	3.31	0.66	2.52
Target Language: Indonesian												
Confucius4-TTS	3.64	0.55	2.79	**2.86**	0.56	2.89	**2.64**	0.72	2.38	3.11	0.67	2.56
OmniVoice	7.62	0.58	**3.00**	11.51	0.68	2.97	3.34	0.77	2.57	**2.51**	0.69	**2.78**
MOSS-TTS	13.07	0.54	2.92	9.33	0.59	**3.01**	8.24	0.74	**2.64**	10.85	0.62	2.75
VoxCPM2	12.27	**0.65**	2.72	10.13	**0.70**	2.71	5.41	**0.78**	2.28	3.36	**0.72**	2.51
Ours	**3.12**	0.44	2.53	3.09	0.58	2.57	3.10	0.70	2.37	5.83	0.66	2.34

they provide noisier realization cues than the curated anchor. This suggests that anchor quality substantially affects linguistic accuracy.

Importantly, the proposed tokenizer does not require a large amount of expressive speech supervision. We train it with approximately 65 hours of emotional speech in addition to 2k hours bilingual read-speech datasets. This relatively small amount of expressive data is sufficient to improve linguistic accuracy and speaker similarity across the evaluated directions.

6 Conclusion

We introduced X-EMO, a disentangled speech tokenization framework for cross-lingual zero-shot expressive TTS. By conditioning the tokenizer decoder on a separate utterance with matched speaker and emotion but different linguistic content, our method encourages the discrete tokens to retain information necessary for linguistic reconstruction while reducing redundant encoding of speaker and emotional attributes. The resulting representation enables linguistic modeling and paralinguistic control to be handled by separate components: a language model predicts the discrete tokens, while a flow-matching acoustic model injects speaker and emotion information from a prompt utterance. Experiments on English, Chinese, Indonesian, and Malay show that the proposed representation supports cross-lingual synthesis with competitive intelligibility, speaker similarity, naturalness, and emotion preservation. Ablation studies further show that an appropriate information cost and expressive tokenizer training improve the learned representation. These results suggest that conditional information bottlenecks provide a simple and data-efficient way to reduce the entanglement of linguistic and paralinguistic information in token-based multilingual TTS.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for collecting relevant literature, code debugging, setting up and troubleshooting the software environments and dependencies required to reproduce baselines. We have not used generative AI tools to develop the overall paper structure or motivation, formulate the key research ideas, or implement the core methods of our proposed approach. Additionally, we used generative AI tools for manuscript language polishing, writing evaluation scripts for running baseline models on our test sets. We have reviewed all AI-assisted work. AI-generated code and evaluation results were manually reviewed and verified by the authors before being used in our experiments and reported results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Elok Anggrayni, Agus Muhamad Hatta, Dhany Arifianto, Joko Sarwono, et al. Inesco dataset: Indonesian expressive speech corpus for emotion speech synthesis. *Data in Brief*, pp. 113058, 2026.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, Marseille, France, 2020. European Language Resources Association.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 818–824, 2024. doi: 10.1109/SLT61566.2024.10832315.

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

Frederik Bous and Axel Roebel. Vasab: The variable size adaptive information bottleneck for disentanglement on speech and singing voice. *arXiv preprint arXiv:2310.03444*, 2023.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pp. 1–5, 2017. doi: 10.1109/ICSDA.2017.8384449.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008. doi: 10.1007/s10579-008-9076-6.

Carlos Busso, Reza Lotfian, K. Sridhar, Ali N. Salman, Wei-Cheng Lin, Lucas Goncalves, S. Parthasarathy, Abinay Reddy Naini, Seong-Gyun Leem, Luz Martinez-Lucas, Huang-Cheng Chou, and Pravin Mote. The msp-podcast corpus. *IEEE Transactions on Affective Computing*, 17:3065–3083, 2025. doi: 10.1109/TAFFC.2026.3678489.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

Dekun Chen, Xueyao Zhang, Yuancheng Wang, Kenan Dai, Li Ma, and Zhizheng Wu. Flexivoice: Enabling flexible style control in zero-shot TTS with natural language instructions. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=F7GmbfyVg9.

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

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 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)*, pp. 6255–6271, 2025.

Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. Large-scale self-supervised speech representation learning for automatic speaker verification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6147–6151. IEEE, 2022b.

Yahuan Cong, Haoyu Zhang, Haopeng Lin, Shichao Liu, Chunfeng Wang, Yi Ren, Xiang Yin, and Zejun Ma. Genertts: Pronunciation disentanglement for timbre and style generalization in cross-lingual text-to-speech. *arXiv preprint arXiv:2306.15304*, 2023.

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

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Xian Shi, Keyu An, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Changfeng Gao, Zhihao Du, and Shiliang Zhang. Differentiable reward optimization for llm based tts system. 2025. URL https://arxiv.org/abs/2507.05911.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, et al. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*, 2023.

Cheng Gong, Chunyu Qiang, Tianrui Wang, Yu Jiang, Yuheng Lu, Ruihao Jing, Xiaoxiao Miao, Xiaolei Zhang, Longbiao Wang, and Jianwu Dang. Perturbation self-supervised representations for cross-lingual emotion tts: Stage-wise modeling of emotion and speaker. *arXiv preprint arXiv:2510.11124*, 2025.

Yitian Gong, Botian Jiang, Yiwei Zhao, Yucheng Yuan, Kuangwei Chen, Yaozhou Jiang, Cheng Chang, Dong Hong, Mingshu Chen, Ruixiao Li, et al. Moss-tts technical report. *arXiv preprint arXiv:2603.18090*, 2026.

Ashishkumar Gudmalwar, Nirmesh Shah, Sai Akarsh, Pankaj Wasnik, and Rajiv Ratn Shah. VECL-TTS: Voice identity and Emotional style controllable Cross-Lingual Text-to-Speech. In *Interspeech 2024*, pp. 3000–3004, 2024. doi: 10.21437/Interspeech.2024-1672.

Yingxu He, Zhuohan Liu, Shuo Sun, Bin Wang, Wenyu Zhang, Xunlong Zou, Nancy F Chen, and Ai Ti Aw. Meralion-audiollm: Bridging audio and language with large language models. *arXiv preprint arXiv:2412.09818*, 2024.

Kevin Heffernan, Artyom Kozhevnikov, Loic Barrault, Alexandre Mourachko, and Holger Schwenk. Aligning speech segments beyond pure semantics. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3626–3635, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.216. URL https://aclanthology.org/2024.findings-acl.216/.

Nina Hosseini-Kivanani and Sandipana Dowerah. Luxemo: Expressive text-to-speech corpus for luxembourgish. *arXiv preprint arXiv:2606.31947*, 2026.

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

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, et al. Ditar: Diffusion transformer autoregressive modeling for speech generation. *arXiv preprint arXiv:2502.03930*, 2025.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiangyang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and sheng zhao. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=dVhrnjZJad.

Yuke Li, Xinfa Zhu, Yi Lei, Hai Li, Junhui Liu, Danming Xie, and Lei Xie. Zero-shot emotion transfer for cross-lingual speech synthesis. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023.

Yunpei Li, Xun Zhou, Jinchao Wang, Lu Wang, Yong Wu, Siyi Zhou, Yiquan Zhou, Yining Wang, Yaogen Yang, Zhetao Hu, et al. Indextts 2.5 technical report. *arXiv preprint arXiv:2601.03888*, 2026.

Shijia Liao, Yuxuan Wang, Songting Liu, Yifan Cheng, Ruoyi Zhang, Tianyu Li, Shidong Li, Yisheng Zheng, Xingwei Liu, Qingzheng Wang, et al. Fish audio s2 technical report. *arXiv preprint arXiv:2603.08823*, 2026.

Qingyu Liu, Yushen Chen, Zhikang Niu, Chunhui Wang, Yunting Yang, Bowen Zhang, Jian Zhao, Pengcheng Zhu, Kai Yu, and Xie Chen. Cross-lingual f5-tts: Towards language-agnostic voice cloning and speech synthesis. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17362–17366. IEEE, 2026.

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

Mozilla Data Collective. Common voice spontaneous speech 4.0 - bahasa malay. https://mozilladatacollective.com/datasets/cmqi2dktu005omf072yc3vzy1, 2026. Accessed: 2026-09-14.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pp. 5210–5219. PMLR, 2019.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*, pp. 7836–7846. PMLR, 2020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 25719–25733, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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

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

Huaxuan Wang, Huimin Wang, Ruiyu Zhang, Yingjie Li, and Yitao Duan. Confucius4-tts: Transcript-free cross-lingual zero-shot tts with a learnable speaker encoder. *arXiv preprint arXiv:2608.11650*, 2026.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. Cam++: A fast and efficient network for speaker verification using context-aware masking. *arXiv preprint arXiv:2303.00332*, 2023b.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. In *International Conference on Learning Representations*, volume 2025, pp. 47127–47150, 2025.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16133–16142. Ieee, 2023.

Rixi Xu, Qingyu Liu, Haitao Li, Yushen Chen, Zhikang Niu, Yunting Yang, Jian Zhao, Ke Li, Berrak Sisman, Qinyuan Cheng, et al. X-voice: Enabling everyone to speak 30 languages via zero-shot cross-lingual voice cloning. *arXiv preprint arXiv:2605.05611*, 2026.

Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, Yaohui Jin, Qingqing Zhang, Pengyuan Zhang, et al. Open source magicdata-ramc: A rich annotated mandarin conversational (ramc) speech dataset. *arXiv preprint arXiv:2203.16844*, 2022.

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

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Guangyan Zhang, Ying Qin, Wenjie Zhang, Jialun Wu, Mei Li, Yutao Gai, Feijun Jiang, and Tan Lee. iemotts: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody and timbre. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1693–1705, 2023a.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech language models. In *International Conference on Learning Representations*, volume 2024, pp. 31798–31818, 2024.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, et al. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In *International Conference on Learning Representations*, volume 2025, pp. 58436–58459, 2025.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*, 2023b.

Zhisheng Zheng, Xiaohang Sun, Zhu Liu, Caren Chen, Rohith Kumar, Manoj Aggarwal, Gerard Medioni, and David Harwath. Ctrlspeech: Coarse-to-fine control for expressive speech synthesis. *arXiv preprint arXiv:2608.08362*, 2026.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 920–924, 2021.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Jiancheng Gui, Jiaheng Wu, Ziyang Wang, Xudong Shen, Runchuan Ye, et al. Voxcpm2 technical report. *arXiv preprint arXiv:2606.06928*, 2026.

Han Zhu, Lingxuan Ye, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhifeng Han, Weiji Zhuang, Long Lin, and Daniel Povey. Omnivoice: Towards omnilingual zero-shot text-to-speech with diffusion language models. *arXiv preprint arXiv:2604.00688*, 2026.

Xinfa Zhu, Yi Lei, Tao Li, Yongmao Zhang, Hongbin Zhou, Heng Lu, and Lei Xie. Metts: Multilingual emotional text-to-speech by cross-speaker and cross-lingual emotion transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1506–1518, 2024.

Tianlun Zuo, Jingbin Hu, Yuke Li, Xinfa Zhu, Hai Li, Ying Yan, Junhui Liu, Danming Xie, and Lei Xie. Xemorag: Cross-lingual emotion transfer with controllable intensity using retrieval-augmented generation. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7. IEEE, 2025.

A Appendix

A.1 Tokenizer Architecture

We use HuBERT-Large as the speech encoder and extract features from its 18th layer. The resulting features are processed by a ConvNeXt-based temporal module (Woo et al., 2023) and projected to the codebook dimension before quantization. We use a codebook size of 8,192 and perform code assignment based on cosine distance.

The decoder is a Transformer that takes the discrete representation together with the conditioning utterance as input. Specifically, the layer-18 HuBERT features of the conditioning utterance are used as a prompt and concatenated with the latent representation along the temporal dimension, with the conditioning features placed before the latent. The resulting sequence is then fed to the decoder, allowing the decoder to directly access information from the conditioning utterance when reconstructing the input speech.

The conditioning utterance is therefore provided to the decoder only as an additional prompt-like sequence and is not passed through the encoder to produce the discrete codes. At inference time, the conditioning utterance is not required for token extraction.

A.2 LANGUAGE MODEL INPUT SEQUENCE FORMAT AND TRAINING DETAILS

Sequence format. We use a unified vocabulary consisting of 8,192 speech tokens and 50,257 Whisper-large-v3 BPE text tokens, structural special tokens, and language tags. Following the text-to-token formulation used in prior work, the input and output are organized into a single sequence, with language tags and text serving as conditioning context and speech tokens as the prediction targets. We use two sequence formats:

single	BOS_IN $\langle l_a \rangle$ text EOS_IN BOS_OUT $\mid$ speech EOS_OUT
code-switch	BOS_IN $\langle l_a \rangle$ text_A $\langle l_b \rangle$ text_B EOS_IN BOS_OUT speech_A $\mid$ speech_B EOS_OUT.

Here, $\langle l_a \rangle$ and $\langle l_b \rangle$ denote language tags, and $\mid$ indicates the beginning of the prediction region. The loss is computed only on the speech tokens and the final EOS_OUT token. The single format is used for standard single-language utterances. The code-switch format is used for utterances containing a language switch, with the corresponding language tags inserted at the language boundary.

Since code-switching data account for only 0.03% of the overall training data, we use the single format for all test-time inputs and do not evaluate the code-switch format. The code-switch format is included primarily to support future extensions to code-switching speech synthesis. We do not make claims about the language model's code-switching capability, nor do we conduct a large-scale evaluation of code-switched generation.

Model and optimization. The language model is implemented as a decoder-only LLaMA (Touvron et al., 2023) with hidden size 1024, 12 layers, 16 attention heads, and an FFN dimension of 4096. With $K = 8192$ speech tokens, the unified vocabulary contains 58,470 tokens. The model has approximately 321M parameters with untied input and output embeddings.

The model is trained for 600k steps with a batch size of 16 using AdamW. The learning rate is 3×10^{-4} with weight decay 0.01. We use 10k warm-up steps followed by cosine decay to 10% of the peak learning rate, with gradient clipping at 1.0. Training is performed in bf16 with a maximum sequence length of 2,048. Text is lowercased before BPE tokenization. Speech tokens are extracted offline using the frozen speech tokenizer. Utterances longer than 35 seconds are discarded.

A.3 FLOW-MATCHING MODEL ARCHITECTURE

The flow-matching model is based on a 16-layer LLaMA backbone with a hidden width of 1024, 16 attention heads, and an FFN dimension of 4096. Since the model predicts the velocity field bidirectionally rather than autoregressively, the causal attention mask is replaced with a padding-only mask.

The conditioning follows a DiT-style design (Peebles & Xie, 2023). The intermediate mel-spectrogram x_t is projected by an MLP, while the speech-token condition is projected by a separate MLP and added to the mel representation. At each transformer layer, the normalization is implemented using adaptive RMSNorm (Zhang & Sennrich, 2019), whose gain is generated from the flow-step embedding. Global conditioning signals are incorporated into the flow-step representation and consequently modulate all transformer layers.

The acoustic target is represented as a 128-band log-mel spectrogram extracted from 24kHz waveforms. We use an FFT size of 1920, a hop size of 480 samples, and a window size of 1920 samples, with the maximum frequency set to 12kHz. The 480-sample hop at 24kHz gives a frame rate of exactly 50Hz, matching the speech-token rate. The mel-spectrograms are normalized using fixed corpus statistics. Waveforms are reconstructed from the generated mel-spectrograms using a frozen pretrained Vocos vocoder (Siuzdak, 2024).

A.4 DATA CLEANING PROCESS

We perform a multi-stage data cleaning procedure to improve the quality and consistency of the speech data. First, we verify the provided transcriptions using two independent ASR systems, Qwen3-ASR (Shi et al., 2026) and MERaLiON-3 (He et al., 2024). For each utterance, we transcribe the speech using both ASR models and compute the WER or CER between the provided

Table 6: Distribution of the training data across languages, measured by the total speech duration. Code-switching utterances are reported separately from the four primary languages.

Language	# Utterances	Hours	Proportion
Chinese	22,638,563	54,798.0	55.01%
English	8,582,253	27,443.7	27.55%
Indonesian	10,578,446	11,305.2	11.35%
Malay	2,582,894	6,047.4	6.07%
Code-switch	20,321	27.2	0.03%
Total	**44,402,477**	**99,621.9**	**100.00%**

transcription and each ASR hypothesis. An utterance is discarded if the WER/CER exceeds 0.3 for either of the two ASR systems.

We then examine the number of speakers of each utterance and remove samples containing multiple speakers or silent segments. Finally, we evaluate the perceptual quality of each speech segment using DNSMOS and discard samples with a DNSMOS score below 2.5. This procedure ensures that the retained data have reliable transcriptions, a single speaker, continuous speech, and sufficient audio quality.

A.5 Training Data Distribution

Table 6 summarizes the language distribution of our training data. The dataset contains 44.4M utterances across four primary languages: Chinese, English, Indonesian, and Malay. Chinese accounts for the largest proportion of the data, while code-switching utterances represent only 0.03% of the total.

A.6 Same-Lingual Emotion Evaluation.

We further evaluate emotion preservation in the same-lingual setting, where the prompt and target utterances are in the same language. The results are presented in Table 7. Overall, our method achieves competitive emotion performance across all three languages. In particular, for English, our method obtains an accuracy of 0.627 and a weighted F1 score of 0.723, which are comparable to the strongest baseline results of 0.647 and 0.727, respectively. It also achieves competitive class-wise F1 scores, including 0.750 for happiness and 0.675 for sadness.

For Chinese, our method achieves the best overall accuracy, weighted F1, and angry-class F1, outperforming the strongest baseline by 1.3, 1.3, and 9.4, respectively. Its happiness and sadness F1 scores remain competitive with the best baseline results. For Indonesian, although our method does not achieve the highest overall accuracy, it obtains a competitive weighted F1 score of 0.445 and the second-best angry-class F1 of 0.712. These results suggest that the proposed representation can preserve emotion-related information even when no cross-lingual transfer is involved, providing an additional sanity check that the emotion information is not lost during the proposed disentanglement and generation process.

A.7 Extended Baseline Comparison

In the main paper, we compare against a representative subset of existing cross-lingual TTS systems. In this appendix, we report all evaluated baselines, including those that do not support all four languages: X-Voice (Xu et al., 2026), CosyVoice3 (Du et al., 2025), Confucius4-TTS (Wang et al., 2026), IndexTTS-2.5 (Li et al., 2026), OmniVoice (Zhu et al., 2026), MOSS-TTS-v1.5 (Gong et al., 2026), F5-TTS (Chen et al., 2025), Fish Audio S2 Pro (Liao et al., 2026), and VoxCPM2 (Zhou et al., 2026), to provide a broader comparison. The results are shown in Table 8. We follow the same evaluation metric as in the main paper and report results for all supported prompt–target language pairs.

Overall, the extended comparison shows similar trends to those observed in the main paper. In particular, no single baseline consistently performs best across intelligibility, speaker similarity, and

Table 7: Emotion preservation results for same-lingual generation, where the prompt and target languages are the same. The best result within each direction and metric is highlighted in bold.

Direction	System	Acc	W-F1	$\mathbf{F1}_{\mathrm{ang}}$	$\mathbf{F1}_{\mathrm{hap}}$	$\mathbf{F1}_{\mathrm{sad}}$
en → en	CosyVoice3	0.580	0.678	0.706	0.723	0.605
	Confucius4-TTS	0.613	0.694	0.755	0.643	**0.683**
	IndexTTS-2.5	**0.647**	**0.727**	**0.756**	**0.760**	0.667
	Ours	0.627	0.723	0.744	0.750	0.675
zh → zh	CosyVoice3	0.460	0.575	0.493	0.660	0.571
	X-Voice	0.380	0.500	0.553	0.535	0.413
	IndexTTS-2.5	0.680	0.769	0.659	**0.929**	**0.718**
	Ours	**0.693**	**0.782**	**0.753**	0.894	0.701
id → id	Confucius4-TTS	0.420	0.431	**0.760**	0.185	0.349
	X-Voice	0.353	0.412	0.638	0.352	0.246
	OmniVoice	**0.493**	**0.582**	0.699	**0.417**	**0.630**
	Ours	0.413	0.445	0.712	0.348	0.276

naturalness. Systems with strong speaker similarity often exhibit substantially higher WER/CER, especially for cross-lingual generation into low-resource languages. For example, VoxCPM2 achieves relatively high speaker similarity across several target languages, but its intelligibility degrades considerably for Malay and Indonesian, with WERs of 6.32 and 12.27 for English-to-Malay and English-to-Indonesian generation, respectively. A similar trade-off can be observed for several other recent baselines.

The difficulty of cross-lingual generation is particularly evident for Malay and Indonesian. While some systems achieve reasonable performance when the prompt and target languages are the same, their intelligibility can deteriorate substantially when the language pair changes. For instance, OmniVoice obtains a speaker similarity of 0.82 for Malay-to-Malay generation, but its WER increases to 6.62 and 7.65 for English-to-Malay and Chinese-to-Malay generation, respectively. In contrast, our model maintains relatively stable intelligibility across prompt languages for several target languages. For Malay, it obtains WERs of 3.21, 3.30, 7.46, and 3.31 for English, Chinese, Malay, and Indonesian prompts, respectively, while for Indonesian it achieves 3.12, 3.09, 3.10, and 5.83.

These results further demonstrate the challenges of multilingual and cross-lingual TTS, particularly for low-resource target languages. Although our model does not achieve the highest speaker similarity or MOS in every setting, it provides a comparatively balanced performance across the three evaluation dimensions while supporting all four target languages. The extended comparison therefore provides additional evidence that the observations in the main paper are not specific to the selected subset of baselines.

Table 8: Extended cross-lingual TTS results on English, Chinese, Malay, and Indonesian. The table follows the same evaluation protocol as Table 5 and additionally includes baselines that do not support all four target languages.

System	Prompt Language											
	English			Chinese			Malay			Indonesian		
	WER/CER	SIM	MOS	WER/CER	SIM	MOS	WER/CER	SIM	MOS	WER/CER	SIM	MOS
Target Language: English												
X-Voice	1.84	0.63	2.98	1.87	0.48	**3.16**	–	–	–	2.46	0.38	3.03
CosyVoice3	2.15	0.69	3.00	2.00	0.53	3.05	–	–	–	–	–	–
Confucius4-TTS	**1.44**	0.70	2.86	**1.80**	0.58	2.85	2.45	0.68	2.37	2.47	0.60	2.66
IndexTTS-2.5	2.66	0.62	2.86	2.33	0.50	3.00	–	–	–	–	–	–
OmniVoice	1.66	0.74	3.02	2.41	**0.67**	2.87	2.49	0.73	2.59	2.06	0.60	2.98
MOSS-TTS-v1.5	2.38	0.69	3.00	3.06	0.59	2.96	5.56	0.69	**2.62**	8.12	0.53	2.87
F5-TTS	1.73	0.68	3.04	1.98	0.58	3.09	–	–	–	–	–	–
Fish Audio S2 Pro	1.69	0.65	**3.10**	2.75	0.59	2.95	–	–	–	2.75	0.46	**3.08**
VoxCPM2	1.61	**0.75**	2.72	2.39	0.66	2.65	4.13	**0.75**	2.29	3.04	**0.68**	2.53
Ours	2.21	0.68	2.81	2.29	0.58	2.86	**2.22**	0.52	2.58	**2.02**	0.43	2.73
Target Language: Chinese												
X-Voice	1.66	0.44	2.66	1.39	0.75	2.82	–	–	–	2.20	0.48	2.77
CosyVoice3	**1.19**	0.44	2.69	1.24	0.78	2.75	–	–	–	–	–	–
Confucius4-TTS	1.23	0.51	2.63	1.01	0.76	2.70	1.83	0.68	2.31	1.80	0.58	2.60
IndexTTS-2.5	2.40	0.45	2.63	1.70	0.74	2.71	–	–	–	–	–	–
OmniVoice	1.46	0.52	2.79	**0.87**	0.78	2.65	7.02	0.74	2.50	**1.52**	0.56	2.84
MOSS-TTS-v1.5	2.41	0.47	2.76	1.30	0.76	2.81	8.90	0.70	**2.53**	6.65	0.52	2.79
F5-TTS	1.97	0.47	2.79	1.62	0.76	**2.85**	–	–	–	–	–	–
Fish Audio S2 Pro	3.94	0.47	**2.88**	0.99	0.74	2.81	–	–	–	3.64	0.49	**2.91**
VoxCPM2	2.41	**0.54**	2.49	1.24	**0.79**	2.45	7.94	**0.76**	2.23	7.98	**0.65**	2.57
Ours	2.51	0.50	2.61	2.56	0.76	2.64	**1.80**	0.68	2.34	1.79	0.54	2.45
Target Language: Malay												
Confucius4-TTS	**3.17**	0.59	2.67	**2.26**	0.61	2.71	**2.24**	0.78	2.41	**2.53**	0.70	2.45
OmniVoice	6.62	0.63	**2.89**	7.65	**0.71**	2.77	2.79	**0.82**	2.58	2.64	0.72	**2.69**
MOSS-TTS-v1.5	4.48	0.50	2.76	5.12	0.59	**2.81**	5.37	0.80	**2.61**	6.02	0.65	2.65
VoxCPM2	6.32	**0.66**	2.53	5.02	0.70	2.51	3.21	**0.82**	2.30	2.72	**0.75**	2.33
Ours	3.21	0.46	2.63	3.30	0.63	2.73	7.46	0.79	2.30	3.31	0.66	2.52
Target Language: Indonesian												
X-Voice	3.33	0.44	2.78	3.34	0.55	2.99	–	–	–	3.34	0.63	2.71
Confucius4-TTS	3.64	0.55	2.79	**2.86**	0.56	2.89	**2.64**	0.72	2.38	3.11	0.67	2.56
OmniVoice	7.62	0.58	3.00	11.51	0.68	2.97	3.34	0.77	2.57	**2.51**	0.69	2.78
MOSS-TTS-v1.5	13.07	0.54	2.92	9.33	0.59	3.01	8.24	0.74	**2.64**	10.85	0.62	2.75
Fish Audio S2 Pro	20.53	0.57	**3.07**	20.52	0.65	**3.02**	–	–	–	6.26	0.60	**2.93**
VoxCPM2	12.27	**0.65**	2.72	10.13	**0.70**	2.71	5.41	**0.78**	2.28	3.36	**0.72**	2.51
Ours	**3.12**	0.44	2.53	3.09	0.58	2.57	3.10	0.70	2.37	5.83	0.66	2.34