CoSTA: Compressed Speech Token Adaptation for Efficient Spoken Dialogue Modeling

Anonymous ACL submission

Abstract

Spoken Dialogue Models (SDMs) typically generate long and redundant discrete speech-token sequences, which limits inference efficiency and the number of interaction turns. A natural solution is to design lower-rate speech tokenizers, but it may introduce a compression-quality trade-off and high replacement cost for an existing SDM. We propose CoSTA, a practical framework that adapts an existing SDM to generate losslessly compressed speech tokens while keeping the original speech tokenizer and speech decoder unchanged. CoSTA applies byte-pair encoding to speech-token sequences, extends the model vocabulary with merged speech tokens, and performs pronunciation and dialogue adaptation. During inference, the generated compressed tokens are deterministically expanded back to the original token space before waveform synthesis. Experiments on bilingual speech-dialogue benchmarks show that CoSTA improves speech-token generation efficiency while preserving response quality, achieving 172.43% and 165.94% relative speed with 66.81% and 70.17% compression rates on Chinese and English benchmarks, respectively. Further analyses show that CoSTA reduces first-packet latency and improves multi-turn sustainability under a limited generation budget.

1 Introduction

Spoken dialogue models (SDMs) enable end-to-end speech interaction by modeling speech input and generating text tokens together with speech tokens, which are then synthesized into waveforms by a speech decoder (Défossez et al., 2024; Zeng et al., 2024; Xu et al., 2025a; KimiTeam et al., 2025; Wu et al., 2025). However, existing SDMs directly model the speech tokens produced by their paired speech tokenizer. These sequences are often overly long, increasing autoregressive generation cost, complicating text-speech alignment, and making long-context speech modeling more difficult (Shen et al., 2024; Li et al., 2024). In real-time spoken dialogue, this inefficiency slows response generation and limits both interaction turns and spoken-response duration under a fixed context or generation budget.

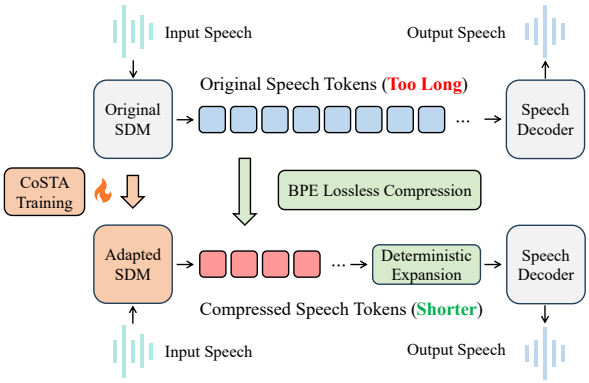

Figure 1: Overview of CoSTA. Existing SDMs directly model the original long speech tokens. CoSTA adapts an existing SDM to generate shorter BPE-compressed speech tokens, which are losslessly expanded before being fed into the unchanged speech decoder.

A natural solution is to design speech tokenizers with lower token rates (Défossez et al., 2024; Zeng et al., 2025; Guo et al., 2024; Wang et al., 2025c; Guo et al., 2026). However, more aggressive tokenization may introduce a compression-quality trade-off, making it more challenging to preserve speech reconstruction or generation quality. Moreover, replacing the tokenizer of an existing SDM is costly, as the speech tokenizer, paired speech decoder, and SDM often need to be retrained or adapted together. This raises a key question: *Can we losslessly compress speech tokens without modifying the speech tokenizer or speech decoder?*

In this work, we exploit the lossless compressibility of the speech-token sequences produced by the original speech tokenizer paired with an existing SDM. These sequences contain repeated local patterns, making them naturally suitable for byte-pair encoding (BPE), a lossless compression algorithm (Gage, 1994) widely used in text language models (Sennrich et al., 2016; Radford et al., 2019).

Each merged BPE token can be deterministically expanded back to original speech tokens, allowing the SDM to model shorter sequences while the original speech decoder still receives its expected token format, as illustrated in Figure 1. To make this compression usable for the existing SDM, we propose CoSTA, short for **Co**mpressed **S**peech **T**oken **A**daptation. Instead of redesigning the speech tokenizer or speech decoder, CoSTA keeps the original speech generation pipeline unchanged and adapts the SDM to model a losslessly compressed speech-token space. Specifically, CoSTA constructs a BPE vocabulary over speech-token sequences, extends the SDM vocabulary with mean-based initialization, and uses progressive pronunciation adaptation followed by dialogue adaptation to learn compressed-token generation. At inference time, it only inserts an extremely lightweight BPE expansion step before waveform synthesis.

Experiments on bilingual speech-dialogue benchmarks show that CoSTA reaches 172.43% and 165.94% relative expanded speech-token generation speed on Chinese and English benchmarks, with compression rates of 66.81% and 70.17%, respectively, while preserving response quality. Further analyses show CoSTA lowers first-packet latency and improves multi-turn sustainability under a limited generation budget.

Our contributions are summarized as follows:

- We show that speech-token sequences produced by the SDM's original speech tokenizer are amenable to BPE compression, enabling compact speech-token modeling while keeping the original speech generation pipeline unchanged.
- We propose CoSTA, a practical adaptation framework that equips an existing SDM with compressed-token generation ability through BPE vocabulary extension with mean-based initialization, staged adaptation, and lightweight inference-time expansion.
- Experiments on bilingual speech-dialogue benchmarks show that CoSTA substantially improves speech-token generation efficiency while preserving response quality. Additional analyses further confirm its benefits in reducing streaming latency and sustaining interaction under constrained generation budgets.

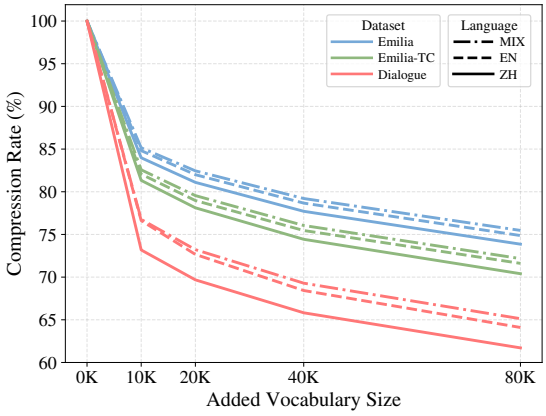

Figure 2: Compression rates with different added BPE vocabulary sizes across Emilia (He et al., 2024), timbre-converted Emilia, and dialogue data under Chinese, English, and mixed-language settings.

2 Speech Tokens Are Losslessly Compressible with BPE

The key observation behind CoSTA is that discrete speech-token sequences contain repeated local patterns, making them naturally suitable for lossless compression. We apply BPE to original speech-token sequences. Let $\mathcal{S}$ be the original speech-token vocabulary and $\mathbf{s} = (s_1, \ldots, s_T)$ be an original speech-token sequence. BPE constructs an extended vocabulary $\mathcal{Z} \supseteq \mathcal{S}$ by merging frequent adjacent token pairs. Let $B(\cdot)$ denote the BPE encoder and $\pi(\cdot)$ denote the deterministic BPE expansion function. The encoder maps $\mathbf{s}$ into a compressed sequence $\mathbf{z} = B(\mathbf{s})$, while the expansion function maps it back to the original speech-token sequence, i.e., $\pi(\mathbf{z}) = \mathbf{s}$. Thus, the expanded sequence remains fully compatible with the original speech decoder.

We measure compactness by the compression rate:

$$\rho(\mathbf{z}) = \frac{|\mathbf{z}|}{|\pi(\mathbf{z})|}, \tag{1}$$

where a lower ρ means fewer autoregressive speech-token steps for representing the same original speech-token sequence consumed by the speech decoder. Figure 2 shows that BPE consistently lowers the compression rate as the extended vocabulary grows across Emilia data, timbre-converted Emilia data, and synthesized dialogue responses, under Chinese, English, and mixed-language settings. Converting speech into a unified timbre further improves compressibility by making token patterns more consistent across utterances. These results show consistent compressibility across data sources and language settings, motivating us to adapt an

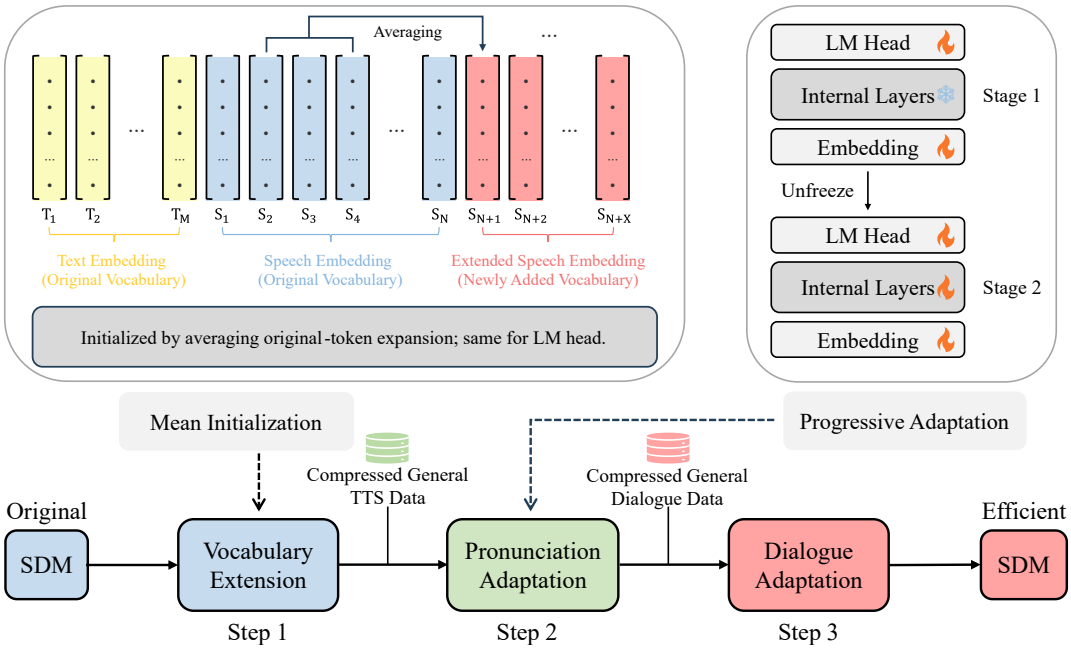

Figure 3: Overview of CoSTA. The framework adapts an existing SDM to generate BPE-compressed speech tokens through mean-based vocabulary initialization, progressive pronunciation adaptation, and dialogue adaptation.

existing SDM to generate BPE-compressed speech tokens directly.

3 Method

In the above analysis, we observe that speech-token sequences produced by the tokenizer paired with an existing SDM are losslessly compressible with BPE. However, existing SDMs are trained on the original speech-token vocabulary and cannot directly model the newly introduced BPE speech tokens. To address this problem, we propose CoSTA, a practical framework that adapts an existing SDM to generate BPE-compressed speech tokens. As Figure 3 shows, it consists of three steps: (1) vocabulary extension with mean-based initialization, (2) pronunciation adaptation on compressed general TTS data, and (3) dialogue adaptation on compressed general speech-to-speech dialogue data. After adaptation, inference only adds a lightweight BPE expansion step before the unchanged speech decoder. We detail each step below.

3.1 Mean-based Initialization for Extended Speech Tokens

To enable the SDM to generate BPE-compressed speech tokens, we extend the original vocabulary $\mathcal{V}$ with newly merged BPE speech tokens:

$$\mathcal{V}' = \mathcal{V} \cup (\mathcal{Z} \setminus \mathcal{S}), \tag{2}$$

where $\mathcal{Z} \setminus \mathcal{S}$ denotes the added BPE speech tokens. Since each added BPE token z has a known expansion $\pi(z) = (s_1, \ldots, s_k)$ into original speech tokens, we initialize its embedding $\mathbf{E}'_z$ and LM head row $\mathbf{W}'_z$ by averaging the corresponding component rows:

$$\mathbf{E}'_z = \tfrac{1}{k}\sum_{i=1}^{k} \mathbf{E}_{s_i}, \quad \mathbf{W}'_z = \tfrac{1}{k}\sum_{i=1}^{k} \mathbf{W}_{s_i}. \tag{3}$$

Original-token parameters remain unchanged. This mean-based initialization provides a more stable starting point than random initialization.

3.2 Pronunciation Adaptation

After vocabulary extension, the model can represent the new tokens but cannot yet reliably generate them. We therefore perform pronunciation adaptation using general TTS data whose target speech tokens are compressed with the learned BPE tokenizer. Following interleaved-generation SDMs such as Step-Audio 2 mini (Wu et al., 2025), the target sequence contains both text tokens and speech tokens. We apply BPE compression only to speech tokens and leave text tokens unchanged.

Each example consists of a text prompt p and a target response sequence $\mathbf{y}$, where $\mathbf{y}$ contains interleaved text tokens and original speech tokens. Let $\tilde{\mathbf{y}} = C(\mathbf{y})$ denote the sequence obtained by applying BPE compression only to the speech tokens of $\mathbf{y}$. We train the model to predict the compressed

target sequence autoregressively:

$$\mathcal{L}_{\text{tts}}(\theta) = -\sum_{(p,\widetilde{\mathbf{y}})\in\mathcal{D}_{\text{tts}}} \sum_{t=1}^{|\widetilde{\mathbf{y}}|} \log P\left(\widetilde{y}_t \mid p, \widetilde{\mathbf{y}}_{<t}\right). \tag{4}$$

To enable the SDM to smoothly acquire the ability to generate more compact speech tokens, we adopt a progressive adaptation strategy inspired by prior work (Li et al., 2025a, 2026). In the first stage, we use 50% of the data to fine-tune only the embedding and LM head layers of the vocabulary-extended model, while keeping the remaining model parameters frozen. This further calibrates the extended vocabulary before broader model adaptation. In the second stage, we use the remaining 50% of the data to jointly fine-tune the intermediate model layers, thereby completing pronunciation adaptation.

3.3 Dialogue Adaptation

Pronunciation adaptation teaches the model to generate compressed speech tokens from text prompts, but it does not fully recover spoken-dialogue ability with the new token vocabulary. We therefore conduct dialogue adaptation on general speech-to-speech dialogue data, where only the LLM parameters are updated while the speech encoder, adaptor, and speech decoder remain frozen. Each dialogue example contains a user's speech query q and a target assistant response $\mathbf{r}$. Similar to pronunciation adaptation, the target response contains interleaved text tokens and original speech tokens. Let $\widetilde{\mathbf{r}} = C(\mathbf{r})$ denote the response after compressing only the speech tokens of $\mathbf{r}$. The objective of dialogue adaptation is:

$$\mathcal{L}_{\text{dlg}}(\theta) = -\sum_{(q,\widetilde{\mathbf{r}})\in\mathcal{D}_{\text{dlg}}} \sum_{t=1}^{|\widetilde{\mathbf{r}}|} \log P\left(\widetilde{r}_t \mid q, \widetilde{\mathbf{r}}_{<t}\right). \tag{5}$$

After this stage, the model acquires the ability to generate compressed speech tokens in speech-to-speech dialogue, and the generated speech tokens can be subsequently expanded by BPE for waveform synthesis.

3.4 Inference with BPE Expansion

As shown in Figure 4, CoSTA does not require any modification to the original speech decoder. Instead, it only inserts a lightweight BPE expansion module after the SDM generates speech tokens. During inference, the adapted SDM first generates an interleaved sequence of text tokens and speech tokens from the extended vocabulary. The generated speech tokens may contain both original tokens and newly introduced BPE tokens. The BPE module leaves original tokens unchanged and expands each merged BPE token into its corresponding sequence of original tokens. The expanded original speech tokens are then fed into the original speech decoder to synthesize the waveform response. Since BPE expansion is deterministic and extremely lightweight, it adds little overhead while reducing the number of autoregressive speech-token steps performed by the SDM.

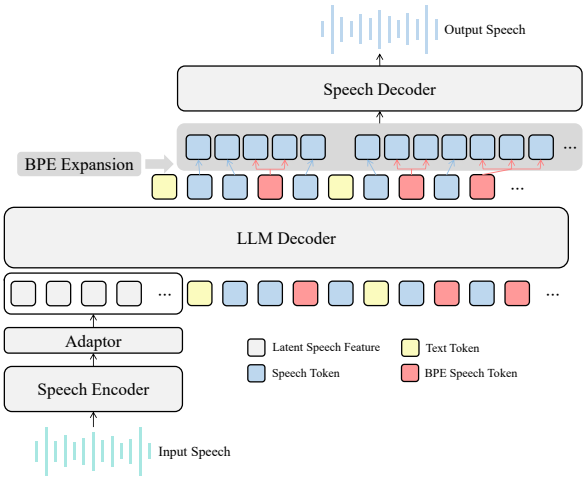

Figure 4: Inference with BPE expansion. The adapted SDM generates BPE-compressed speech tokens, which are losslessly expanded and then synthesized by the unchanged speech decoder.

4 Experiments

4.1 Experimental Setup

Implementation. We use Step-Audio 2 mini (Wu et al., 2025), an 8B-parameter model, as the base SDM due to its relatively complete open-source support, simple and efficient architecture, and competitive spoken-dialogue performance. For pronunciation adaptation, we sample 300 hours of Chinese speech and 300 hours of English speech from Emilia (He et al., 2024) and convert them to a unified timbre with Seed-VC (Liu, 2024). For dialogue adaptation, we follow the OpenS2S (Wang et al., 2025a) data construction pipeline to synthesize general Chinese and English speech-dialogue data, with target speech generated by CosyVoice 2 (Du et al., 2024b) using a consistent voice. We train the BPE tokenizer on dialogue speech-token sequences and use 80K additional BPE speech tokens in the main experiments. Figure 2 shows that

Model	AlpacaEval-zh		APE-zh		Claude-zh		LCSTS-zh		MLC-zh		Wildchat-zh		Avg		
	$\rho_g \downarrow$	Speed↑	$\rho_g \downarrow$	Speed↑	$\rho_g \downarrow$	Speed↑	$\rho_g \downarrow$	Speed↑	$\rho_g \downarrow$	Speed↑	$\rho_g \downarrow$	Speed↑	$\rho_g \downarrow$	Speed↑	Relative↑
Step-Audio 2 mini	100.00%	57.97	100.00%	60.88	100.00%	56.39	100.00%	57.05	100.00%	65.22	100.00%	57.40	100.00%	59.15	100.00%
Vanilla-FT-ZH	100.00%	57.84	100.00%	60.40	100.00%	54.99	100.00%	55.54	100.00%	63.47	100.00%	59.10	100.00%	58.56	99.00%
Vanilla-FT-MIX	100.00%	59.01	100.00%	59.65	100.00%	56.93	100.00%	56.03	100.00%	61.57	100.00%	59.08	100.00%	58.71	99.26%
CoSTA-ZH	**67.50%**	**103.03**	**64.63%**	101.68	**68.52%**	**101.20**	**69.40%**	**94.46**	**62.70%**	**110.31**	**68.13%**	**101.29**	**66.81%**	**101.99**	**172.43%**
CoSTA-MIX	69.12%	98.44	64.68%	**102.34**	70.58%	92.54	71.06%	91.37	63.03%	106.32	70.21%	93.22	68.11%	97.37	164.62%
Model	**AlpacaEval**		**Gsm8kEval**		**CommonEval**		**Summary**		**MLC**		**WildchatEval**		**Avg**		
	$\rho_g \downarrow$	Speed↑	$\rho_g \downarrow$	Speed↑	$\rho_g \downarrow$	Speed↑	$\rho_g \downarrow$	Speed↑	$\rho_g \downarrow$	Speed↑	$\rho_g \downarrow$	Speed↑	$\rho_g \downarrow$	Speed↑	Relative↑
Step-Audio 2 mini	100.00%	57.93	100.00%	60.03	100.00%	60.62	100.00%	58.74	100.00%	61.80	100.00%	59.00	100.00%	59.69	100.00%
Vanilla-FT-EN	100.00%	56.73	100.00%	58.14	100.00%	56.96	100.00%	56.16	100.00%	61.36	100.00%	56.67	100.00%	57.67	96.61%
Vanilla-FT-MIX	100.00%	57.43	100.00%	59.27	100.00%	57.29	100.00%	57.66	100.00%	61.90	100.00%	56.43	100.00%	58.33	97.72%
CoSTA-EN	**73.97%**	**87.60**	65.60%	**102.04**	**71.60%**	**98.41**	**72.09%**	**95.21**	64.71%	115.58	**73.06%**	**95.46**	**70.17%**	**99.05**	**165.94%**
CoSTA-MIX	75.45%	81.17	**65.55%**	97.92	72.34%	88.91	73.94%	83.29	**64.30%**	**123.95**	74.56%	87.80	71.02%	93.84	157.21%

Table 1: Efficiency comparison on Chinese and English speech-to-speech benchmarks. ρ_g denotes the generated-token compression rate, Speed denotes the expanded speech-token generation speed, and Relative denotes the average speed relative to Step-Audio 2 mini within each language setting.

larger added BPE vocabularies consistently reduce the compression rate, with gradually diminishing gains at larger sizes. We choose 80K as a high-compression setting. The vocabulary extension is lightweight: it only adds embedding and LM head rows, which account for about 6% of the original model parameters. We train CoSTA-ZH, CoSTA-EN, and CoSTA-MIX for Chinese, English, and bilingual adaptation, respectively. More details are provided in Appendices A and B.1.

Baselines. We compare CoSTA with Step-Audio 2 mini (Wu et al., 2025) and vanilla fine-tuning baselines without BPE compression. Step-Audio 2 mini is the original base model, while Vanilla-FT-ZH, Vanilla-FT-EN, and Vanilla-FT-MIX are fine-tuned on the same Chinese, English, and bilingual adaptation data as their CoSTA counterparts. This controls for the effect of training data scale and isolates the impact of BPE-compressed speech-token adaptation. We also report Qwen2.5-Omni (Xu et al., 2025a), Kimi-Audio (KimiTeam et al., 2025), and GLM-4-Voice (Zeng et al., 2024) as external quality references.

Benchmarks. We evaluate speech-to-speech dialogue ability on 12 datasets from URO-Bench (Yan et al., 2025), covering aligned Chinese and English settings. The evaluation includes summarization (LCSTS-zh and Summary), practical mathematical problems (APE-zh and Gsm8kEval), math, logic, and common sense (MLC-zh and MLC), authentic open-ended dialogue (AlpacaEval-zh, AlpacaEval, Claude-zh, and CommonEval), and real-world conversation (Wildchat-zh and WildchatEval). These benchmarks better reflect realistic spoken dialogue scenarios, making them well aligned with our goal of improving the efficiency and usability of speech-to-speech SDMs.

Metrics. We evaluate *efficiency* with three metrics. First, we report the **generated-token compression rate** ρ_g, computed using the compression-rate definition in Eq. 1 on the model-generated speech-token sequence $\widehat{\mathbf{z}}$. A lower ρ_g indicates that the model generates a more compact speech-token sequence. Second, we report the **expanded speech-token generation speed**:

$$\text{Speed} = \frac{|\pi(\widehat{\mathbf{z}})|}{T_{\text{decode}}}, \tag{6}$$

which measures how many original speech tokens the model effectively produces per second after BPE expansion. Third, we report **relative speed**, denoted as **Relative** in the tables, defined as the ratio between a model's average expanded speech-token generation speed and that of Step-Audio 2 mini under the same language setting. This metric normalizes speed improvements against the original SDM. For output *quality*, we follow the URO-Bench evaluation protocol and report **G-Eval**, **UTMOS** (Saeki et al., 2022), and **WER/CER** for response-content quality, speech quality, and text-speech alignment, respectively. Evaluation details are provided in Appendix B.2.

4.2 Main Results

Efficiency. Table 1 reports the generated-token compression rate ρ_g, expanded speech-token generation speed, and relative speed over Step-Audio 2 mini under the vLLM streaming inference setting. Fine-tuning without compression does not improve efficiency: Vanilla-FT-ZH, Vanilla-FT-EN, and Vanilla-FT-MIX keep $\rho_g = 100\%$ and

Model	AlpacaEval-zh	APE-zh	Claude-zh	LCSTS-zh	MLC-zh	Wildchat-zh	Avg		
							G-Eval↑	ASR-CER↓	UTMOS↑
Qwen2.5-Omni	74.10	62.11	78.86	69.75	82.53	78.35	74.28	2.93%	2.95
Kimi-Audio	77.05	39.12	80.24	76.13	80.92	78.08	71.93	4.88%	2.52
GLM-4-Voice	84.76	23.16	83.54	76.36	78.85	84.41	71.85	3.77%	2.94
Step-Audio 2 mini	80.18	21.93	72.40	79.16	78.16	**83.59**	69.24	7.49%	3.62
Vanilla-FT-ZH	79.77	21.40	80.35	78.82	77.47	81.47	69.88	4.96%	**3.77**
Vanilla-FT-MIX	81.27	20.53	79.28	79.10	79.31	82.10	70.26	7.02%	**3.77**
CoSTA-ZH	**81.36**	**24.56**	**86.49**	**80.56**	**79.54**	81.09	**72.27**	**1.51%**	3.75
CoSTA-MIX	80.95	**24.56**	85.05	78.99	77.24	82.39	71.53	1.65%	3.75
Model	**AlpacaEval**	**Gsm8kEval**	**CommonEval**	**Summary**	**MLC**	**WildchatEval**	**Avg**		
							G-Eval↑	**ASR-WER↓**	**UTMOS↑**
Qwen2.5-Omni	76.78	54.30	72.00	80.06	77.40	76.71	72.88	1.73%	4.28
Kimi-Audio	62.01	0.57	50.77	82.15	40.87	66.40	50.46	34.92%	2.86
GLM-4-Voice	80.87	29.32	66.77	92.15	69.30	81.26	69.95	11.55%	4.16
Step-Audio 2 mini	76.31	15.92	52.53	88.14	72.69	76.75	63.73	6.49%	4.53
Vanilla-FT-EN	73.74	16.67	72.03	86.10	72.50	78.47	66.59	10.16%	4.46
Vanilla-FT-MIX	76.18	11.05	72.63	85.37	68.55	78.93	65.45	12.00%	4.49
CoSTA-EN	81.34	15.52	**75.87**	89.21	**74.01**	**83.84**	**69.97**	**2.50%**	**4.54**
CoSTA-MIX	**83.32**	**17.53**	73.47	**90.00**	71.19	82.52	69.67	2.64%	**4.54**

Table 2: Quality comparison on Chinese and English speech-to-speech benchmarks. Benchmark columns report G-Eval scores, while the Avg columns report G-Eval, ASR-CER/WER, and UTMOS for response-content quality, text-speech alignment, and speech quality, respectively. **Bold** and underline mark the best and second-best results among Step-Audio 2 mini variants only; external models are quality references.

achieve nearly the same average speed as the base model. In contrast, CoSTA substantially reduces the generated-token compression rate across both languages and diverse benchmark types. CoSTA-ZH lowers the average ρ_g to 66.81% and achieves 172.43% relative speed over Step-Audio 2 mini on Chinese benchmarks, while CoSTA-EN reaches 70.17% compression rate and 165.94% relative speed on English benchmarks. The bilingual CoSTA-MIX variants show the same trend, indicating that the efficiency benefit is robust across language settings. Table 1 shows a clear trend: lower compression rates generally yield higher speech-generation speed. Overall, these results show that making speech-token sequences more compact effectively reduces autoregressive generation cost and improves speech-token generation efficiency.

Quality. Table 2 evaluates whether the efficiency gains of CoSTA come at the cost of response-content quality, speech quality, or text-speech alignment. Overall, CoSTA preserves or improves the quality of the base SDM while substantially improving efficiency. This shows that lossless speech-token compression does not merely accelerate generation, but can also make the token-generation process easier to model. On Chinese benchmarks, CoSTA-ZH improves response-content quality over both Step-Audio 2 mini and vanilla fine-tuned baselines, while greatly reducing ASR-CER. This suggests that compressed speech-token modeling not only accelerates generation but also improves text-speech alignment. Meanwhile, UTMOS remains comparable to vanilla fine-tuning, showing that the unchanged speech decoder can still synthesize natural speech after BPE expansion. On English benchmarks, CoSTA-EN shows the same trend: it improves average G-Eval, reduces ASR-WER, and maintains high UTMOS. The bilingual CoSTA-MIX variants achieve comparable quality in both languages, suggesting that the adaptation strategy generalizes beyond monolingual settings. These results highlight that CoSTA improves the efficiency-quality trade-off of an existing SDM without replacing its speech tokenizer or speech decoder. CoSTA is therefore complementary to architectural improvements as well as model and data scaling, offering a practical way to make existing spoken dialogue models more efficient while preserving their spoken-dialogue quality.

4.3 Ablation Study

Effect of each component. Table 3 shows that the components of CoSTA contribute complementary effects. Timbre conversion improves token-pattern consistency by reducing speaker variation, leading to better compressed-token learning and response quality. Progressive adaptation further lowers the compression rate of the generated speech tokens, indicating that a smoother adaptation process helps the model learn more compact token generation. Pronunciation adaptation is necessary for teaching the model to generate the new compressed speech tokens reliably. Removing it weakens compression and text-speech alignment, show-

Variant	$\rho_g \downarrow$	ASR-CER↓	UTMOS↑	G-Eval↑
Step-Audio 2 mini	100.00%	7.49%	3.62	69.24
Vanilla-FT-ZH	100.00%	4.96%	**3.77**	69.88
CoSTA-ZH	**66.73%**	1.51%	3.75	**72.27**
w/o Timbre Conversion	67.01%	1.52%	3.75	70.93
w/o Progressive Adaptation	67.75%	**1.46%**	3.75	72.19
w/o Pronunciation Adaptation	69.56%	1.74%	3.74	72.12
w/o Dialogue Adaptation	81.05%	2.68%	3.74	24.35

Table 3: Component ablation on Chinese benchmarks. ρ_g, ASR-CER, UTMOS, and G-Eval evaluate compression efficiency, text-speech alignment, speech quality, and response-content quality, respectively.

ing that the model benefits from a dedicated TTS-style adaptation stage before dialogue training. Dialogue adaptation is even more crucial for recovering spoken-dialogue ability: without it, response quality drops sharply despite partial compressed-token generation. We further compare initialization strategies and learning rates in the first phase of pronunciation adaptation, including select-based initialization, which copies the embedding and output-projection row of a randomly selected token from each merged token's original-token expansion. As shown in Figures 5 and 6, mean-based initialization with a learning rate of 1.6×10^{-4} achieves faster loss reduction and more stable gradients, indicating a better starting point for adapting the expanded vocabulary.

Effect of expanded vocabulary size. Table 4 shows that a larger extended BPE vocabulary helps the model generate more compact speech-token sequences. Increasing the number of added tokens from 10K to 80K reduces ρ_g from 76.39% to 66.73%, while response quality remains stable or slightly improves. This indicates that extending the BPE vocabulary strengthens lossless compression without sacrificing speech quality or spoken-dialogue performance.

4.4 Analyses

We further analyze CoSTA from two practical perspectives: streaming first-packet latency and interaction sustainability under a limited generation budget. Additional analyses on data scaling and case studies are provided in Appendices C.3 and D.

CoSTA reduces first-packet latency. Table 5 reports the first-packet latency in streaming deployment for both Chinese and English settings, measured from the moment the SDM receives the input speech. This latency consists of LLM-side latency and speech-decoder latency. Different from the model-side speed in Table 1, first-packet latency reflects the full streaming pipeline before the first audio packet is produced. CoSTA consistently reduces total first-packet latency compared with both the original model and the vanilla fine-tuned baseline. The improvement mainly comes from the LLM side, where compressed speech-token generation shortens the time needed to produce enough tokens for the first speech packet. In contrast, speech-decoder latency changes only mildly because CoSTA keeps the original speech decoder unchanged. This shows CoSTA primarily accelerates the autoregressive generation stage, and its end-to-end latency benefit is naturally bounded by the unchanged downstream speech synthesis cost.

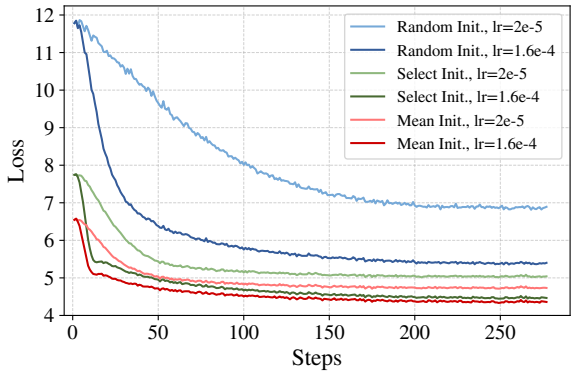

Figure 5: Training loss during the first stage of pronunciation adaptation. Mean-based initialization with a larger learning rate achieves faster and more stable convergence.

Model	Size	$\rho_g \downarrow$	ASR-CER↓	UTMOS↑	G-Eval↑
Step-Audio 2 mini	–	100.00%	7.49%	3.62	69.24
Vanilla-FT-ZH	–	100.00%	4.96%	**3.77**	69.88
CoSTA-ZH	10K	76.39%	1.49%	3.76	71.31
	20K	73.53%	**1.34%**	3.76	71.98
	40K	70.35%	1.52%	3.75	72.02
	80K	**66.73%**	1.51%	3.75	**72.27**

Table 4: Effect of added BPE vocabulary size on Chinese benchmarks. ρ_g, ASR-CER, UTMOS, and G-Eval measure generated-token compression, text-speech alignment, speech quality, and response-content quality, respectively.

CoSTA improves interaction sustainability under a limited budget. Table 6 evaluates interaction sustainability on AlpacaEval-zh and AlpacaEval under a limited generation budget. We construct stress-test sessions by grouping every 10 speech queries, yielding 14 Chinese sessions and 19 English sessions. This setting is not intended as a conventional multi-turn dialogue quality evaluation, but rather tests whether the model can sustain the required interleaved text-speech generation pattern under budget constraints. We report the average number of effective dialogue turns, the number

Model	$\rho_g \downarrow$	Speed↑	Latency (ms)↓		
			LLM	Speech Decoder	Total
Step-Audio 2 mini	100.00%	59.15	220.44	337.46	557.90
Vanilla-FT-ZH	100.00%	58.56	219.59	340.51	560.10
CoSTA-ZH	**66.81%**	**101.99**	**177.53**	**328.06**	**505.59**
Step-Audio 2 mini	100.00%	59.69	218.15	344.44	562.60
Vanilla-FT-EN	100.00%	57.67	220.82	339.52	560.34
CoSTA-EN	**70.17%**	**99.05**	**169.20**	**339.17**	**508.36**

Table 5: First-packet latency on Chinese and English speech-to-speech benchmarks. The upper and lower blocks report Chinese and English results, respectively.

Model	# Turns↑	Gen Tokens	Final Tokens	Duration (s)	$\rho_g \downarrow$
Step-Audio 2 mini	2.14	2208.21	2208.21	88.33	100.00%
Vanilla-FT-ZH	3.71	2440.50	2440.50	97.62	100.00%
CoSTA-ZH	**7.21**	**2500.79**	**3766.86**	**150.67**	**66.39%**
Step-Audio 2 mini	2.58	2218.74	2218.74	88.75	100.00%
Vanilla-FT-EN	1.84	1894.79	1894.79	75.79	100.00%
CoSTA-EN	**5.79**	**2608.53**	**3650.79**	**146.03**	**71.45%**

Table 6: Multi-turn interaction under a limited generation budget. The upper and lower blocks report Chinese and English results, respectively. Gen/Final Tokens denote speech tokens before/after BPE expansion and Duration denotes the generated speech duration.

of speech tokens before and after BPE expansion within these effective turns, and the corresponding generated speech duration. An effective turn is counted only when the model output still follows the original interleaved token generation pattern. The results show that CoSTA sustains more effective turns and longer generated speech duration in both languages, even though its adaptation data does not contain multi-turn dialogue tasks. This demonstrates a practical benefit of compressed generation: the model can represent more spoken content and maintain interaction for longer under a limited generation budget.

5 Related Works

5.1 Spoken Dialogue Models

Existing SDMs directly model speech input and spoken responses in a unified framework, and mainly differ in how speech generation is coupled with the language model. Think-Talk architectures use separate Transformer backbones for text and speech generation (Fang et al., 2025a; Wang et al., 2025b; Fang et al., 2025b; Wang et al., 2025a; Xu et al., 2025a,b; Team, 2026; Cui et al., 2026; Défossez et al., 2024; Team et al., 2025a); parallel-generation architectures use one Transformer backbone with multiple output heads (KimiTeam et al., 2025; Chen et al., 2025; Xie and Wu, 2024a,b; Team et al., 2025b); and interleaved-generation architectures use a single Transformer to model interleaved text-speech generation (Zhang et al., 2023; Li et al., 2025b; Zeng et al., 2024; Wu et al., 2025). While these works advance SDM architectures and interaction ability, they still directly model the original speech-token sequences produced by their paired speech tokenizers. CoSTA instead focuses on the redundancy of these sequences and adapts an existing SDM to generate losslessly compressed speech tokens without changing the original speech generation pipeline.

5.2 Discrete Speech Tokens

Discrete speech tokens are widely used in speech generation and spoken dialogue modeling, including acoustic tokens from codec models (Zeghidour et al., 2022; Défossez et al., 2023; Kumar et al., 2023; Ji et al., 2025; Zhang et al., 2024) and semantic tokens from discriminative (Du et al., 2024a,b, 2025) or self-supervised speech models (Baevski et al., 2020a,b; Hsu et al., 2021; Chen et al., 2022; Chung et al., 2021). Existing SDMs usually generate these original speech-token sequences directly, making their length and redundancy a major efficiency bottleneck. A natural solution is to design lower-token-rate speech tokenizers (Défossez et al., 2024; Zeng et al., 2025; Guo et al., 2024; Wang et al., 2025c; Guo et al., 2026), but this may introduce compression-quality trade-offs and require costly joint adaptation of the tokenizer, speech decoder, and SDM. BPE-style compression has been explored for discrete speech units in ASR (Ren et al., 2022; Chang et al., 2023, 2024), speech continuation (Shen et al., 2024), and TTS (Zhu et al., 2023; Li et al., 2024), revealing reusable local patterns in speech-token sequences. However, these studies mainly target task-specific speech modeling rather than enabling existing SDMs to efficiently generate speech in interactive speech-to-speech dialogue. In contrast, CoSTA enables an existing SDM to generate compressed BPE speech tokens while keeping the original tokenizer-decoder pipeline unchanged.

6 Conclusion

We presented CoSTA, which adapts an existing SDM to generate losslessly compressed speech tokens without changing the original speech generation pipeline. By combining BPE-based compression with vocabulary extension, and staged adaptation, CoSTA reduces autoregressive speech-token generation cost. Experiments show that CoSTA improves speech-token generation efficiency while preserving response quality, reduces streaming latency, and enhances interaction sustainability under constrained generation budgets.

Limitations

This work has several limitations and possible extensions. First, our dialogue-adaptation data is constructed from synthesized speech-to-speech examples. While such data provides controllable and consistent speech responses for studying speech-token compression, it is also a practical and widely adopted choice for training spoken dialogue models, as large-scale, high-quality real speech-to-speech dialogue data is difficult to collect. Future work may further explore how synthetic and real conversational speech data complement each other. Second, we adopt BPE as a deterministic, lossless, and easily integrated compression method. Future work may investigate other lossless speech-token segmentation strategies and compare their behavior in spoken dialogue modeling. Finally, our evaluation focuses on general speech-to-speech dialogue and efficiency. Future work may extend the analysis to finer-grained scenarios, such as emotional dialogue, personalized speaking styles, and diverse acoustic environments.

References

Alexei Baevski, Steffen Schneider, and Michael Auli. 2020a. vq-wav2vec: Self-supervised learning of discrete speech representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xuankai Chang, Brian Yan, Kwanghee Choi, Jee-Weon Jung, Yichen Lu, Soumi Maiti, Roshan S. Sharma, Jiatong Shi, Jinchuan Tian, Shinji Watanabe, Yuya Fujita, Takashi Maekaku, Pengcheng Guo, Yao-Fei Cheng, Pavel Denisov, Kohei Saijo, and Hsiu-Hsuan Wang. 2024. Exploring speech recognition, translation, and understanding with discrete speech units: A comparative study. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 11481–11485. IEEE.

Xuankai Chang, Brian Yan, Yuya Fujita, Takashi Maekaku, and Shinji Watanabe. 2023. Exploration of efficient end-to-end ASR using discretized input from self-supervised learning. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 1399–1403. ISCA.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518.

Wenxi Chen, Ziyang Ma, Ruiqi Yan, Yuzhe Liang, Xiquan Li, Ruiyang Xu, Zhikang Niu, Yanqiao Zhu, Yifan Yang, Zhanxun Liu, Kai Yu, Yuxuan Hu, Jinyu Li, Yan Lu, Shujie Liu, and Xie Chen. 2025. Slam-omni: Timbre-controllable voice interaction system with single-stage training. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 2262–2282. Association for Computational Linguistics.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*, pages 244–250. IEEE.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, and 17 others. 2026. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *Preprint*, arXiv:2604.27393.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2023. High fidelity neural audio compression. *Trans. Mach. Learn. Res.*, 2023.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *Preprint*, arXiv:2407.05407.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, and 3 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *Preprint*, arXiv:2505.17589.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang,

Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *Preprint*, arXiv:2412.10117.

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. *Preprint*, arXiv:2410.00037.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025a. Llama-omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. 2025b. Llama-omni 2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 18617–18629. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2063–2067. ISCA.

Haohan Guo, Fenglong Xie, Kun Xie, Dongchao Yang, Dake Guo, Xixin Wu, and Helen Meng. 2024. Socodec: A semantic-ordered multi-stream speech codec for efficient language model based text-to-speech synthesis. In *IEEE Spoken Language Technology Workshop, SLT 2024, Macao, December 2-5, 2024*, pages 645–651. IEEE.

Yiwei Guo, Zhihan Li, Hankun Wang, Bohan Li, Chongtian Shao, Hanglei Zhang, Chenpeng Du, Xie Chen, Shujie Liu, and Kai Yu. 2026. Recent advances in discrete speech tokens: A review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(4):4184–4204.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *IEEE Spoken Language Technology Workshop, SLT 2024, Macao, December 2-5, 2024*, pages 885–890. 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 Trans. Audio Speech Lang. Process.*, 29:3451–3460.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. 2025. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, Jiyan Yang, Jongsoo Park, Alexander Heinecke, Evangelos Georganas, Sudarshan Srinivasan, Abhisek Kundu, Misha Smelyanskiy, Bharat Kaul, and Pradeep Dubey. 2019. A study of bfloat16 for deep learning training. *Preprint*, arXiv:1905.12322.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Bohan Li, Feiyu Shen, Yiwei Guo, Shuai Wang, Xie Chen, and Kai Yu. 2024. On the effectiveness of acoustic BPE in decoder-only TTS. In *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*. ISCA.

Chong Li, Yingzhuo Deng, Wen Yang, Jiajun Zhang, and Chengqing Zong. 2026. Tokalign++: Advancing vocabulary adaptation via better token alignment. *Preprint*, arXiv:2605.13429.

Chong Li, Jiajun Zhang, and Chengqing Zong. 2025a. Tokalign: Efficient vocabulary adaptation via token alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 4109–4126. Association for Computational Linguistics.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, Jianhua Xu, Haoze Sun, Zenan Zhou, and Weipeng Chen. 2025b. Baichuan-audio: A unified framework for end-to-end speech interaction. *Preprint*, arXiv:2502.17239.

Songting Liu. 2024. Zero-shot voice conversion with diffusion transformers. *Preprint*, arXiv:2411.09943.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

OpenAI. 2024. GPT-4o mini: Advancing Cost-Efficient Intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2026-05-25.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 28492–28518. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020, Virtual Event / Atlanta, Georgia, USA, November 9-19, 2020*, page 20. IEEE/ACM.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.

Shuo Ren, Shujie Liu, Yu Wu, Long Zhou, and Furu Wei. 2022. Speech pre-training with acoustic piece. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2648–2652. ISCA.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: utokyo-sarulab system for voicemos challenge 2022. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 4521–4525. ISCA.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Feiyu Shen, Yiwei Guo, Chenpeng Du, Xie Chen, and Kai Yu. 2024. Acoustic BPE for speech generation with discrete tokens. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 11746–11750. IEEE.

Core Team, Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, Xin Zhang, Xingchen Song, Yihan Yan, Yongzhe He, Cici, Bowen Shen, Chengxuan Zhu, Chong Ma, and 81 others. 2025a. Mimo-audio: Audio language models are few-shot learners. *Preprint*, arXiv:2512.23808.

Meituan LongCat Team, Bairui Wang, Bayan, Bin Xiao, Bo Zhang, Bolin Rong, Borun Chen, Chang Wan, Chao Zhang, Chen Huang, Chen Chen, Chen Chen, Chengxu Yang, Chengzuo Yang, Cong Han, Dandan Peng, Delian Ruan, Detai Xin, Disong Wang, and 114 others. 2025b. Longcat-flash-omni technical report. *Preprint*, arXiv:2511.00279.

Qwen Team. 2026. Qwen3.5-omni technical report. *Preprint*, arXiv:2604.15804.

Chen Wang, Tianyu Peng, Wen Yang, Yinan Bai, Guangfu Wang, Jun Lin, Lanpeng Jia, Lingxiang Wu, Jinqiao Wang, Chengqing Zong, and Jiajun Zhang. 2025a. Opens2s: Advancing fully open-source end-to-end empathetic large speech language model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025 - System Demonstrations, Suzhou, China, November 4-9, 2025*, pages 906–917. Association for Computational Linguistics.

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2025b. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Yuancheng Wang, Dekun Chen, Xueyao Zhang, Junan Zhang, Jiaqi Li, and Zhizheng Wu. 2025c. Tadicodec: Text-aware diffusion speech tokenizer for speech language modeling. *Preprint*, arXiv:2508.16790.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, and 90 others. 2025. Step-audio 2 technical report. *Preprint*, arXiv:2507.16632.

Zhifei Xie and Changqiao Wu. 2024a. Mini-omni: Language models can hear, talk while thinking in streaming. *Preprint*, arXiv:2408.16725.

Zhifei Xie and Changqiao Wu. 2024b. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *Preprint*, arXiv:2410.11190.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. 2025. Urobench: Towards comprehensive evaluation for end-to-end spoken dialogue models. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 17211–17242. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. Soundstream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:495–507.

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. *Preprint*, arXiv:2412.02612.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Lei Zhang, Shengmin Jiang, Yuxiao Dong, and Jie Tang. 2025. Scaling speech-text pre-training with synthetic interleaved data. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 15757–15773. Association for Computational Linguistics.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Xinfa Zhu, Yuanjun Lv, Yi Lei, Tao Li, Wendi He, Hongbin Zhou, Heng Lu, and Lei Xie. 2023. Vec-tok speech: speech vectorization and tokenization for neural speech generation. *Preprint*, arXiv:2310.07246.

A Implementation Details and Hyperparameters

This section reports the implementation details and hyperparameters used in CoSTA. Unless otherwise specified, all models are initialized from Step-Audio 2 mini (Wu et al., 2025). The original speech tokenizer and speech decoder are kept unchanged throughout all experiments. The original speech-token vocabulary contains 6,656 tokens. We train the BPE tokenizer using the HuggingFace `tokenizers` library[1] on speech-token sequences extracted from the general speech-to-speech dialogue corpus, and extend the original vocabulary with 80K additional BPE speech tokens. Since the BPE trainer does not directly operate on integer-valued speech-token sequences, we map each original speech-token ID to a unique Unicode character before BPE training. During BPE expansion, generated BPE-token IDs are first converted into their corresponding Unicode strings, and BPE expansion is performed in the Unicode space; the resulting BPE-expanded Unicode sequence is then mapped back to the corresponding original speech-token IDs. This conversion is deterministic and does not change the underlying speech-token sequence. All adaptation and inference experiments are conducted on NVIDIA A100 80GB GPUs. For all adaptation stages, we use the AdamW optimizer (Loshchilov and Hutter, 2019), train with bfloat16 mixed precision (Micikevicius et al., 2018; Kalamkar et al., 2019), and use DeepSpeed (Rasley et al., 2020) with ZeRO-1 optimizer-state partitioning (Rajbhandari et al., 2020) for distributed training.

A.1 Pronunciation Adaptation

Pronunciation adaptation is conducted on compressed general TTS data. Following the progressive adaptation strategy, we split the pronunciation-adaptation data into two halves. In the first stage, only the embedding and LM head layers of the vocabulary-extended model are updated. We use a relatively large learning rate of 1.6×10^{-4} in this stage, which leads to faster loss reduction and more stable optimization as shown in Figures 5 and 6. In the second stage, intermediate model layers are further unfrozen and trained with a smaller learning rate of 2.0×10^{-5} to complete pronunciation adaptation. The hyperparameters are shown in Table 7.

[1] https://github.com/huggingface/tokenizers

Hyperparameter	Value
Data split	50% / 50%
Stage-1 trainable params	Embedding, LM head
Stage-2 trainable params	+ Intermediate layers
Optimizer	AdamW
Stage-1 learning rate	1.6×10^{-4}
Stage-2 learning rate	2.0×10^{-5}
Scheduler	Cosine
Warmup ratio	0.03
Global batch size	256
Precision	bfloat16
Gradient clipping norm	1.0
Epochs	3
GPUs	8

Table 7: Hyperparameters used during pronunciation adaptation.

Hyperparameter	Value
Trainable params	All LLM parameters
Optimizer	AdamW
Learning rate	2.0×10^{-5}
Scheduler	Cosine
Warmup ratio	0.03
Global batch size	256
Precision	bfloat16
Gradient clipping norm	1.0
Epochs	3
GPUs	8

Table 8: Hyperparameters used during dialogue adaptation.

A.2 Dialogue Adaptation

After pronunciation adaptation, we perform dialogue adaptation on compressed general speech-to-speech dialogue data. The target responses contain interleaved text and speech tokens, where only speech tokens are BPE-compressed and text tokens remain unchanged. The compressed speech tokens are arranged in the response sequence according to the original interleaving ratio between text tokens and speech tokens used by the base SDM. The hyperparameters are shown in Table 8.

A.3 Inference

During inference, the adapted SDM generates an interleaved sequence of text tokens and speech tokens from the extended vocabulary. The generated speech tokens may include both original speech tokens and newly added BPE speech tokens. We apply BPE expansion only to the generated speech-token stream: original speech tokens are kept unchanged, while each merged BPE token is deterministically expanded into its original speech-token sequence. The expanded speech tokens are then

Hyperparameter	Value
Decoding mode	Greedy
Repetition penalty	1.2
Max new tokens	2048
BPE expansion	Enabled
Expansion target	Speech-token stream only
Text-token stream	Unchanged
Speech decoder	Original decoder
Hardware	1 NVIDIA A100 GPU

Table 9: Hyperparameters used during inference.

fed into the unchanged speech decoder for waveform synthesis. The inference hyperparameters are shown in Table 9.

B Training Data and Benchmarks

This section provides additional details about the training data and evaluation benchmarks used in our experiments.

B.1 Training Data

Pronunciation-adaptation data. For pronunciation adaptation, we use Emilia (He et al., 2024), a large-scale multilingual speech generation dataset constructed from in-the-wild speech. Emilia contains diverse speech data across multiple languages and speaking styles, making it suitable for training speech generation models. In this work, we sample 300 hours of Chinese speech and 300 hours of English speech from Emilia. To reduce speaker variation and improve speech-token pattern consistency, we convert the sampled utterances into a unified timbre using Seed-VC (Liu, 2024). The converted speech is then tokenized by the original speech tokenizer of Step-Audio 2 mini (Wu et al., 2025), and the resulting speech tokens are compressed with the learned BPE tokenizer. Text tokens remain unchanged. The compressed speech tokens are arranged according to the original text-speech interleaving pattern used by the base SDM.

Dialogue-adaptation data. For dialogue adaptation, we follow the general-purpose bilingual speech-to-speech instruction-pair construction pipeline of OpenS2S (Wang et al., 2025a). Specifically, we first collect general text instructions from InstructS2S-200K (Fang et al., 2025a) and translate the English instructions into Chinese using Qwen3-32B (Yang et al., 2025), yielding bilingual text instructions. We then select seed audio with neutral emotion labels and convert these text instructions into user speech queries. Following the speech response generation process in OpenS2S, Qwen3-32B is used to generate dialogue-appropriate text responses, and CosyVoice 2 (Du et al., 2024b) synthesizes the corresponding assistant speech with a consistent reference voice. Following this pipeline, we initially construct 50K Chinese and 50K English general speech-to-speech instruction examples, each containing a user's speech query and an assistant response. We further filter out examples whose query or response speech exceeds one minute, resulting in 49,664 Chinese and 49,408 English dialogue examples. The statistics of the training data are summarized in Table 10. During training, only speech tokens in the target response are BPE-compressed, while text tokens remain unchanged. The compressed speech tokens are arranged according to the original text-speech interleaving pattern used by the base SDM.

Data Type	Language	Scale
Text-to-speech	Chinese	300h
	English	300h
Speech dialogue	Chinese	49,664 examples
	English	49,408 examples

Table 10: Statistics of training data.

Dataset	#Samples	Task / Aspect	Eval. Mode
AlpacaEval-zh	147	Authentic, open-ended dialogue	Open
AlpacaEval	199	Authentic, open-ended dialogue	Open
APE-zh	190	Practical mathematical problems	QA
Gsm8kEval	582	Practical mathematical problems	QA
Claude-zh	222	Authentic, open-ended dialogue	Open
CommonEval	200	Authentic, open-ended dialogue	Open
LCSTS-zh	119	Summarization	Semi-open
Summary	118	Summarization	Semi-open
MLC-zh	145	Math, logic, and common sense	QA
MLC	177	Math, logic, and common sense	QA
Wildchat-zh	299	Real-world conversation	Open
WildchatEval	349	Real-world conversation	Open

Table 11: URO-Bench datasets used in our evaluation. The selected datasets cover understanding, reasoning, and oral conversation in Chinese and English speech-to-speech settings.

B.2 Evaluation Benchmarks

We evaluate models on URO-Bench (Yan et al., 2025), a comprehensive benchmark for end-to-end spoken dialogue models in speech-to-speech scenarios. URO-Bench covers three major evaluation dimensions: Understanding, Reasoning, and Oral Conversation, and assesses not only response-content quality but also speech quality and speech-text alignment. In our experiments, we use 12 aligned Chinese and English datasets from URO-

Bench, as summarized in Table 11. These benchmarks better reflect realistic spoken dialogue scenarios, making them well aligned with our goal of improving the efficiency and usability of speech-to-speech SDMs.

Following the URO-Bench evaluation protocol, we evaluate output quality using its prompts and scoring procedures. Specifically, we report the following metrics: (1) **G-Eval** measures response-content quality by transcribing the generated speech with an ASR model and prompting GPT-4o mini (OpenAI, 2024) to score the transcription. We use Whisper-large-v3 (Radford et al., 2023) for English transcription and Paraformer-zh (Gao et al., 2022) for Chinese transcription. (2) **UTMOS** (Saeki et al., 2022) estimates speech quality with a predicted mean opinion score that reflects clarity, naturalness, and fluency. (3) **WER/CER** evaluates the alignment between the speech and text outputs of SDMs by computing the Word Error Rate (WER) or Character Error Rate (CER) between the generated text and the transcription of the generated speech.

URO-Bench uses different evaluation prompts according to the task format. For open-ended tasks, including AlpacaEval-zh, AlpacaEval, Claude-zh, CommonEval, Wildchat-zh, and WildchatEval, we use `prompt_open`. For semi-open tasks with reference answers, including LCSTS-zh and Summary, we use `prompt_semi_open`. For question-answering tasks with deterministic references, including APE-zh, Gsm8kEval, MLC-zh, and MLC, we use `prompt_qa`. These prompts are fed into GPT-4o mini together with the user instruction, the model response, and reference answer when available, to obtain the final G-Eval score. The complete prompts are provided below.

prompt_open

I need your help to evaluate the performance of several models in the speech interaction scenario. The models will receive a speech input from the user, which they need to understand and respond to with a speech output.
Your task is to rate the model's responses based on the provided user input transcription [Instruction] and the model's output transcription [Response].
Please evaluate the response on a scale of 1 to 5:
1 point: The response is largely irrelevant, incorrect, or fails to address the user's query. It may be off-topic or provide incorrect information.
2 points: The response is somewhat relevant but lacks accuracy or completeness. It may only partially answer the user's question or include extraneous information.
3 points: The response is relevant and mostly accurate, but it may lack conciseness or include unnecessary details that do not contribute to the main point.
4 points: The response is relevant, accurate, and concise, providing a clear answer to the user's question without unnecessary elaboration.
5 points: The response is exceptionally relevant, accurate, and to the point. It directly addresses the user's query in a highly effective and efficient manner, providing exactly the information needed.
Below are the transcription of user's instruction and model's response:
[Instruction]
{question}
[Response]
{answer}
After evaluating, please output the score only without anything else. You do not need to provide any explanations.

prompt_semi_open

I need your help to evaluate the performance of several models in the speech interaction scenario. The models will receive a speech input from the user, which they need to understand and respond to with a speech output.
Your task is to rate the model's responses based on the provided user input transcription [Instruction], the model's output transcription [Response] and some suggested answers [Reference]. The model's response does not necessarily have to be identical to the suggested answers, as long as it aligns with the question and is reasonable.
Please evaluate the response on a scale of 1 to 5:
1 point: The response is largely irrelevant, incorrect, or fails to address the user's query. It may be off-topic or provide incorrect information. The response does not align with the question in any meaningful way.
2 points: The response is somewhat relevant but

lacks accuracy, completeness, or coherence. It may partially address the query but introduces unnecessary information or deviates from the core issue. The response may not align well with the suggested answer but still provides some value.
3 points: The response is relevant and mostly accurate, but may lack conciseness or clarity. It addresses the question reasonably, but there might be slight deviations in approach or content. While it may not strictly align with the suggested answer, it still effectively addresses the core of the query.
4 points: The response is relevant, accurate, and concise. It provides a clear answer to the user's question and avoids unnecessary details. While it may not exactly mirror the suggested answer, it effectively addresses the user's query in a logical and well-reasoned manner.
5 points: The response is exceptionally relevant, accurate, and concise. It directly addresses the user's query in the most efficient manner, providing exactly the information needed. The response may differ from the suggested answer in phrasing or approach but still aligns perfectly with the intent of the query, demonstrating a high level of reasoning and clarity.
Below are the transcription of user's instruction, model's response and the reference answer:
[Instruction]
{question}
[Response]
{answer}
[Reference]
{reference}
After evaluating, please output the score only without anything else. You do not need to provide any explanations.

prompt_qa

I need your help to evaluate the performance of several models in the speech interaction scenario. The models will receive a speech input from the user, which they need to understand and respond to with a speech output.
Your task is to rate the model's responses based on the provided user input transcription [Question], the model's output transcription [Response] and the correct answer [Reference].
Below are the transcription of user's instruction, model's response and the reference answer:
[Question]
{question}
[Response]
{answer}
[Reference]
{reference}
Is the model's response correct based on the question and reference answer?
Please only output a single "Yes" or "No". Do not output anything else.

C Additional Results

This section provides additional experimental results that complement the main paper. We first report the detailed per-dataset quality results, then provide the gradient-norm curves for different initialization strategies, and finally present the data-scaling analysis for pronunciation-adaptation data and dialogue-adaptation data.

C.1 Detailed Main Results

In the main paper, we report per-dataset G-Eval scores and average ASR-CER/WER and UTMOS results for Chinese and English URO-Bench datasets. To provide a more complete comparison, Tables 13 and 14 further report the detailed values of all metrics on each dataset, including G-Eval, ASR-CER/WER, and UTMOS. These results show that CoSTA consistently improves or maintains response-content quality, speech-text alignment, and speech naturalness across different task types.

C.2 Gradient Norm of Initialization Strategies

In Section 4, we compare different initialization strategies and learning rates during the first phase of pronunciation adaptation. Figure 6 reports the corresponding gradient-norm curves. Compared with random initialization and select initialization, mean-based initialization leads to a more stable optimization process. In particular, mean-based initialization with a learning rate of 1.6×10^{-4} achieves faster loss reduction while keeping the gradient norm more controlled, indicating that averaging the original-token components provides a better starting point for adapting the newly added BPE speech tokens.

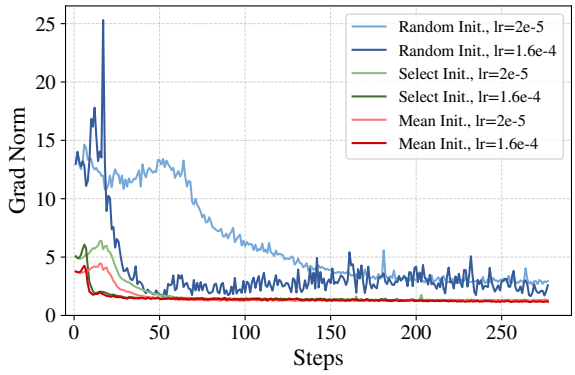

Figure 6: Gradient norm during the first stage of pronunciation adaptation. Mean-based initialization yields more stable optimization under a larger learning rate.

Name	License
Step-Audio 2 mini	Apache 2.0 license
Emilia	CC BY-NC-4.0 license
Seed-VC	GPL-3.0 license
OpenS2S	Apache 2.0 license
CosyVoice 2	Apache 2.0 license
URO-Bench	MIT license
HuggingFace tokenizers	Apache 2.0 license
Transformers	Apache 2.0 license
vLLM	Apache 2.0 license
DeepSpeed	Apache 2.0 license
Whisper	MIT license
Paraformer / FunASR	Model License Agreement
GPT-4o-mini API	OpenAI Service Terms

Table 12: Licenses of scientific artifacts used in this work.

C.3 Data Scaling Analysis

We further analyze how the amount of adaptation data affects CoSTA. These scaling experiments were conducted in an early setting without progressive adaptation, and were mainly evaluated on four Chinese URO-Bench datasets: AlpacaEval-zh, APE-zh, Claude-zh, and MLC-zh. We study two scaling settings: varying the amount of pronunciation-adaptation data while keeping the dialogue-adaptation data fixed, and varying the amount of dialogue-adaptation data while keeping the pronunciation-adaptation data fixed.

Scaling pronunciation-adaptation data. Table 15 reports the results when varying the amount of TTS data used for pronunciation adaptation, while fixing the dialogue-adaptation data to 50K examples. Increasing the amount of TTS data generally improves compressed-token learning, especially in terms of compression rate and speech-text alignment. However, the improvement becomes less pronounced as the TTS data size continues to increase. This suggests that pronunciation adaptation benefits from sufficient TTS data, while the marginal gain becomes smaller beyond a certain scale. Based on this trade-off, we use 300h TTS data as the default setting, which provides reliable compressed speech-token generation while keeping the adaptation cost moderate.

Scaling dialogue-adaptation data. Table 16 reports the results when varying the amount of speech-to-speech dialogue data, while fixing the pronunciation-adaptation data to 300h. The results show that using 50K dialogue examples provides a strong balance between compression behavior and response quality. Therefore, we adopt 300h TTS data and 50K dialogue examples as the experimental configuration.

D Case Study

We present representative case studies selected from the English benchmarks AlpacaEval and WildchatEval. Case Study 1 and Case Study 2 show short factual and simple commonsense QA examples, where CoSTA-EN produces concise responses with much fewer BPE speech tokens. Case Study 3 illustrates a sentence-rewriting example, and Case Study 4 presents a more open-ended dialogue example. Across these cases, CoSTA-EN preserves response quality while generating shorter compressed speech-token sequences, which are deterministically expanded back to the original token space before waveform synthesis. These examples qualitatively demonstrate the efficiency advantage of compressed speech-token generation without changing the original speech decoder.

E Licenses of Scientific Artifacts

We follow and report the licenses of the scientific artifacts used in this work in Table 12.

Model	AlpacaEval-zh			APE-zh			Claude-zh			Avg		
	CER↓	UTMOS↑	G-Eval↑	CER↓	UTMOS↑	G-Eval↑	CER↓	UTMOS↑	G-Eval↑	CER↓	UTMOS↑	G-Eval↑
Qwen2.5-Omni	2.09%	2.93	74.10	8.50%	3.02	62.11	2.47%	2.73	78.86	2.93%	2.95	74.28
Kimi-Audio	4.71%	2.52	77.05	8.29%	2.59	39.12	11.12%	2.08	80.24	4.88%	2.52	71.93
GLM-4-Voice	3.65%	2.86	84.76	5.51%	3.14	23.16	7.13%	2.59	83.54	3.77%	2.94	71.85
Step-Audio 2 mini	8.48%	3.58	80.18	4.35%	3.87	21.93	20.28%	3.33	72.40	7.49%	3.62	69.24
Vanilla-FT-ZH	5.11%	3.76	79.77	0.95%	4.06	21.40	19.51%	3.41	80.35	4.96%	3.77	69.88
Vanilla-FT-MIX	6.32%	3.74	81.27	1.00%	4.12	20.53	27.73%	3.39	79.28	7.02%	3.77	70.26
CoSTA-ZH	1.92%	3.70	81.36	1.37%	4.04	24.56	2.62%	3.43	86.49	1.51%	3.75	72.27
CoSTA-MIX	1.96%	3.70	80.95	2.33%	4.07	24.56	2.55%	3.39	85.05	1.65%	3.75	71.53
Model	**LCSTS-zh**			**MLC-zh**			**Wildchat-zh**			**Avg**		
	CER↓	UTMOS↑	G-Eval↑	CER↓	UTMOS↑	G-Eval↑	CER↓	UTMOS↑	G-Eval↑	CER↓	UTMOS↑	G-Eval↑
Qwen2.5-Omni	1.86%	3.04	69.75	1.31%	3.07	82.53	1.37%	2.88	78.35	2.93%	2.95	74.28
Kimi-Audio	2.42%	2.55	76.13	0.75%	2.83	80.92	1.96%	2.55	78.08	4.88%	2.52	71.93
GLM-4-Voice	1.80%	2.92	76.36	1.74%	3.24	78.85	2.78%	2.86	84.41	3.77%	2.94	71.85
Step-Audio 2 mini	1.28%	3.64	79.16	1.18%	3.85	78.16	9.39%	3.48	83.59	7.49%	3.62	69.24
Vanilla-FT-ZH	0.91%	3.71	78.82	1.23%	3.99	77.47	2.03%	3.70	81.47	4.96%	3.77	69.88
Vanilla-FT-MIX	3.83%	3.67	79.10	0.80%	3.98	79.31	2.44%	3.73	82.10	7.02%	3.77	70.26
CoSTA-ZH	1.18%	3.70	80.56	1.06%	3.96	79.54	0.90%	3.71	81.09	1.51%	3.75	72.27
CoSTA-MIX	1.16%	3.70	78.99	0.77%	3.96	77.24	1.15%	3.65	82.39	1.65%	3.75	71.53

Table 13: Detailed quality results on Chinese speech-to-speech benchmarks. Each benchmark reports CER, UTMOS, and G-Eval for text-speech alignment, speech quality, and response-content quality, respectively.

Model	AlpacaEval			Gsm8kEval			CommonEval			Avg		
	WER↓	UTMOS↑	G-Eval↑	WER↓	UTMOS↑	G-Eval↑	WER↓	UTMOS↑	G-Eval↑	WER↓	UTMOS↑	G-Eval↑
Qwen2.5-Omni	2.05%	4.22	76.78	4.25%	4.32	54.30	1.48%	4.27	72.00	1.73%	4.28	72.88
Kimi-Audio	20.77%	2.53	62.01	99.08%	2.77	0.57	37.62%	2.95	50.77	34.92%	2.86	50.46
GLM-4-Voice	13.92%	4.00	80.87	8.29%	4.25	29.32	26.28%	4.10	66.77	11.55%	4.16	69.95
Step-Audio 2 mini	9.82%	4.51	76.31	17.07%	4.53	15.92	1.64%	4.53	52.53	6.49%	4.53	63.73
Vanilla-FT-EN	23.82%	4.35	73.74	4.15%	4.53	16.67	9.73%	4.47	72.03	10.16%	4.46	66.59
Vanilla-FT-MIX	27.61%	4.43	76.18	3.63%	4.54	11.05	9.89%	4.51	72.63	12.00%	4.49	65.45
CoSTA-EN	3.08%	4.54	81.34	3.01%	4.54	15.52	2.80%	4.55	75.87	2.50%	4.54	69.97
CoSTA-MIX	3.19%	4.54	83.32	3.26%	4.55	17.53	2.61%	4.55	73.47	2.64%	4.54	69.67
Model	**Summary**			**MLC**			**WildchatEval**			**Avg**		
	WER↓	UTMOS↑	G-Eval↑	WER↓	UTMOS↑	G-Eval↑	WER↓	UTMOS↑	G-Eval↑	WER↓	UTMOS↑	G-Eval↑
Qwen2.5-Omni	0.38%	4.34	80.06	1.02%	4.30	77.40	1.20%	4.24	76.71	1.73%	4.28	72.88
Kimi-Audio	3.43%	3.14	82.15	27.72%	3.17	40.87	20.91%	2.58	66.40	34.92%	2.86	50.46
GLM-4-Voice	1.39%	4.26	92.15	10.79%	4.23	69.30	8.61%	4.11	81.26	11.55%	4.16	69.95
Step-Audio 2 mini	0.28%	4.55	88.14	3.23%	4.53	72.69	6.86%	4.52	76.75	6.49%	4.53	63.73
Vanilla-FT-EN	4.18%	4.51	86.10	3.63%	4.53	72.50	15.45%	4.41	78.47	10.16%	4.46	66.59
Vanilla-FT-MIX	10.09%	4.49	85.37	3.96%	4.52	68.55	16.85%	4.46	78.93	12.00%	4.49	65.45
CoSTA-EN	1.34%	4.55	89.21	2.00%	4.53	74.01	2.76%	4.54	83.84	2.50%	4.54	69.97
CoSTA-MIX	1.25%	4.55	90.00	2.07%	4.53	71.19	3.45%	4.54	82.52	2.64%	4.54	69.67

Table 14: Detailed quality results on English speech-to-speech benchmarks. Each benchmark reports WER, UTMOS, and G-Eval for text-speech alignment, speech quality, and response-content quality, respectively.

Model	TTS Data	AlpacaEval-zh				APE-zh				Avg			
		$\rho_g \downarrow$	ASR-CER↓	UTMOS↑	G-Eval↑	$\rho_g \downarrow$	ASR-CER↓	UTMOS↑	G-Eval↑	$\rho_g \downarrow$	ASR-CER↓	UTMOS↑	G-Eval↑
Step-Audio 2 mini	–	100.00%	8.48%	3.54	79.95	100.00%	4.28%	3.87	22.11	100.00%	8.57%	3.66	63.34
CoSTA-ZH	500h	68.17%	1.62%	3.70	80.86	65.32%	1.11%	4.06	21.05	**66.48%**	1.52%	3.79	66.35
	400h	68.73%	1.69%	3.70	82.40	65.88%	1.48%	4.05	25.26	67.35%	1.62%	3.79	67.71
	300h	68.54%	1.91%	3.72	81.59	65.09%	1.02%	4.06	25.09	66.83%	1.66%	3.78	67.75
	200h	69.29%	2.15%	3.68	82.36	65.69%	0.83%	4.07	26.84	67.24%	1.71%	3.78	**67.99**
	100h	68.78%	1.75%	3.73	82.86	65.50%	1.55%	4.03	22.11	67.39%	1.80%	3.78	66.88
Model	**TTS Data**	**Claude-zh**				**MLC-zh**				**Avg**			
		$\rho_g \downarrow$	ASR-CER↓	UTMOS↑	G-Eval↑	$\rho_g \downarrow$	ASR-CER↓	UTMOS↑	G-Eval↑	$\rho_g \downarrow$	ASR-CER↓	UTMOS↑	G-Eval↑
Step-Audio 2 mini	–	100.00%	20.21%	3.36	73.15	100.00%	1.30%	3.85	78.16	100.00%	8.57%	3.66	63.34
CoSTA-ZH	500h	69.35%	2.39%	3.41	84.86	63.09%	0.95%	3.99	78.62	**66.48%**	1.52%	3.79	66.35
	400h	70.07%	2.42%	3.42	85.47	64.70%	0.88%	3.99	77.70	67.35%	1.62%	3.79	67.71
	300h	69.77%	2.98%	3.39	84.77	63.93%	0.71%	3.93	79.54	66.83%	1.66%	3.78	67.75
	200h	70.12%	2.84%	3.40	85.53	63.85%	1.02%	3.97	77.24	67.24%	1.71%	3.78	**67.99**
	100h	70.38%	2.69%	3.39	85.77	64.90%	1.22%	3.95	76.78	67.39%	1.80%	3.78	66.88

Table 15: Effect of pronunciation-adaptation data scale on Chinese benchmarks. Dialogue-adaptation data is fixed to 50K examples. ρ_g denotes the generated-token compression rate, while ASR-CER, UTMOS, and G-Eval evaluate text-speech alignment, speech quality, and response-content quality, respectively.

Model	Dialogue Data	AlpacaEval-zh				APE-zh				Avg			
		$\rho_g \downarrow$	ASR-CER↓	UTMOS↑	G-Eval↑	$\rho_g \downarrow$	ASR-CER↓	UTMOS↑	G-Eval↑	$\rho_g \downarrow$	ASR-CER↓	UTMOS↑	G-Eval↑
Step-Audio 2 mini	–	100.00%	8.48%	3.54	79.95	100.00%	4.28%	3.87	22.11	100.00%	8.57%	3.66	63.34
CoSTA-ZH	50K	68.54%	1.91%	3.72	81.59	65.09%	1.02%	4.06	25.09	**66.83%**	1.66%	3.78	**67.75**
	40K	69.24%	1.79%	3.71	78.59	65.61%	1.36%	4.04	22.81	67.28%	1.72%	3.78	66.02
	30K	68.82%	1.90%	3.73	79.23	66.04%	1.00%	4.04	18.07	67.50%	1.80%	3.78	64.95
	20K	68.32%	2.02%	3.70	80.45	66.20%	0.95%	4.06	25.44	67.49%	1.97%	3.79	66.84
	10K	68.82%	1.31%	3.75	82.59	68.10%	0.67%	4.02	24.56	68.19%	2.09%	3.80	67.25
Model	**Dialogue Data**	**Claude-zh**				**MLC-zh**				**Avg**			
		$\rho_g \downarrow$	ASR-CER↓	UTMOS↑	G-Eval↑	$\rho_g \downarrow$	ASR-CER↓	UTMOS↑	G-Eval↑	$\rho_g \downarrow$	ASR-CER↓	UTMOS↑	G-Eval↑
Step-Audio 2 mini	–	100.00%	20.21%	3.36	73.15	100.00%	1.30%	3.85	78.16	100.00%	8.57%	3.66	63.34
CoSTA-ZH	50K	69.77%	2.98%	3.39	84.77	63.93%	0.71%	3.93	79.54	**66.83%**	1.66%	3.78	**67.75**
	40K	69.99%	2.68%	3.41	86.13	64.28%	1.07%	3.97	76.55	67.28%	1.72%	3.78	66.02
	30K	70.25%	2.90%	3.38	84.80	64.88%	1.39%	3.98	77.70	67.50%	1.80%	3.78	64.95
	20K	70.19%	2.93%	3.45	86.05	65.25%	1.97%	3.95	75.40	67.49%	1.97%	3.79	66.84
	10K	70.30%	2.50%	3.50	84.59	65.55%	3.86%	3.95	77.24	68.19%	2.09%	3.80	67.25

Table 16: Effect of dialogue-adaptation data scale on Chinese benchmarks. Pronunciation-adaptation data is fixed to 300h. ρ_g denotes the generated-token compression rate, while ASR-CER, UTMOS, and G-Eval evaluate text-speech alignment, speech quality, and response-content quality, respectively.

Case Study 1: Short Factual QA

Input: *What breed of dog is the smallest?*

Step-Audio 2 mini

Text Content: The Chihuahua is considered one of the smallest dog breeds, though there are other tiny dogs like the Yorkshire Terrier and Pomeranian that can be even smaller than some Chihuahuas.

Speech Tokens:

[1475, 1978, 4218, 4137, 2031, 2004, 72, 876, 2490, 5322, 3111, 3887, 2183, 1456, 710, 3404, 5591, 5026, 5830, 5092, 5023, 4699, 1490, 1763, 4769, 4450, 5644, 1623, 168, 12, 3908, 3910, 5645, 4673, 2217, 4526, 138, 3151, 3872, 1686, 1854, 2185, 2915, 5024, 3744, 443, 4275, 2085, 315, 65, 4672, 4916, 4915, 4914, 2124, 6501, 2849, 5597, 4958, 5074, 4909, 4492, 4411, 4753, 3459, 1950, 2112, 60, 2333, 2759, 2681, 4949, 5078, 1128, 2076, 2166, 2160, 4367, 1685, 962, 395, 557, 79, 1833, 6379, 6374, 5401, 27, 1458, 1707, 3648, 3645, 1701, 1704, 1947, 3891, 3894, 1950, 1959, 1914, 306, 2243, 2810, 2912, 5828, 1044, 4194, 4442, 2384, 3107, 4106, 1892, 1865, 4079, 1919, 2003, 1976, 6320, 4779, 543, 4528, 6059, 1687, 303, 2214, 6075, 5599, 5189, 2285, 3666, 2726, 2753, 1599, 66, 461, 494, 1223, 1220, 704, 1209, 81, 5482, 6372, 6475, 2702, 3002, 475, 2139, 63, 80, 1294, 3480, 3884, 5090, 5806, 4114, 732, 6048, 5319, 3141, 3875, 1680, 2247, 2274, 2922, 2366, 5294, 6293, 4115, 3158, 3401, 959, 1928, 1676, 1520, 737, 22, 3667, 2175, 1431, 5589, 5598, 5671, 2845, 4311, 2135, 2171, 2180, 1652, 4553, 5047, 2373, 2938, 2699, 539, 125, 5, 1476, 1480, 1723, 1716, 2040, 1272, 2216, 197, 1059, 2355, 25, 6178, 2172, 969, 4939, 2005, 2014, 5019, 4858, 4347, 2315, 3664, 1726, 6377, 5402, 4918, 1170, 4314, 662, 5678, 5025, 5062, 2750, 380, 1688, 884, 1281, 74, 20, 3913, 6376, 5648, 4671, 4619, 658, 4072, 1041, 4593, 2382, 1700, 2186, 1457, 575, 4133, 6077, 2894, 4372, 719, 515, 29, 794, 4428, 5726, 5645, 2484, 1702, 6405, 4218, 4218, 4218]

Length: **267** Speech Tokens

Vanilla-FT-EN

Text Content: The smallest breed of dog is the Chihuahua.

Speech Tokens:

[3704, 4299, 6486, 6405, 4218, 1950, 1560, 801, 2246, 4428, 6377, 1035, 2130, 2125, 4949, 4945, 5029, 5059, 5985, 4409, 4733, 4444, 6454, 4191, 2232, 1188, 2166, 2169, 4367, 3872, 2420, 5102, 231, 70, 1760, 3404, 4868, 2894, 6534, 2112, 60, 2423, 5051, 4946, 494, 491, 515, 1125, 84, 168, 1707, 5838, 5109, 2922, 979, 1951, 386, 4841, 4454, 5409, 2085, 72, 2327, 150, 2733, 5326, 3111, 3806, 2102, 2185, 710, 6077, 6157, 5050, 3643, 707, 3161, 2918, 5832, 1701, 4299, 4299, 4299, 4299, 6486, 6486, 6486, 4299, 4299, 4299, 6486, 6486, 6405, 4218, 4218, 4218, 4299, 6159]

Length: **98** Speech Tokens

CoSTA-EN

Text Content: The smallest dog breed is the chihuahua.

Speech Tokens:

[3704, 7192, 6725, 6676, 1959, 3831, 45, 4442, 6792, 1035, 14382, 3410, 9992, 5068, 6148, 4459, 4498, 4429, 6979, 2209, 795, 7411, 353, 4946, 13225, 4770, 975, 6705, 2169, 9941, 5300, 2627, 11198, 484, 786, 7553, 17346, 4454, 8462, 3225, 140, 2336, 1356, 17640, 1072, 626, 1456, 3643, 3488, 25896, 5062, 5830, 2786, 17267, 974, 2927, 5832, 6669, 13230, 22981, 9416]

Expanded Speech Tokens:

[3704, 4299, 6486, 6405, 4218, 4218, 4218, 6405, 6405, 1959, 3831, 45, 4442, 6373, 6377, 1035, 2130, 2125, 3410, 4865, 4942, 5068, 6148, 4459, 4498, 4429, 2139, 4431, 2209, 795, 2139, 66, 353, 4946, 5675, 4942, 4770, 975, 4299, 2166, 2169, 4358, 6050, 5300, 2627, 638, 566, 484, 786, 1763, 1951, 2573, 4760, 4454, 4429, 4272, 3225, 140, 2336, 1356, 5326, 3840, 1072, 626, 1456, 3643, 3488, 6157, 5508, 5062, 5830, 2786, 3161, 974, 974, 2927, 5832, 4299, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 6405, 6405, 6405, 6405, 6405, 6405, 4218, 4218, 4218, 4218, 4218, 4299, 6078]

Length: **61** BPE Speech Tokens / **101** Expanded Speech Tokens

Compression Rate: **60.40%**

Case Study 2: Simple Commonsense QA

Input: *What color is the sky?*

Step-Audio 2 mini

Text Content: The sky is blue.

Speech Tokens:

[1499, 4218, 6486, 4299, 2112, 1029, 74, 2261, 6455, 6374, 1890, 1788, 183, 2270, 5591, 5348, 5186, 4543, 4821, 5011, 1978, 1034, 4517, 5179, 6382, 1044, 2175, 2166, 721, 6475, 4690, 4613, 3155, 884, 2099, 2102, 2186, 2149, 4299, 6486, 6405, 4218]

Length: **42** Speech Tokens

Vanilla-FT-EN

Text Content: The sky is blue.

Speech Tokens:

[1517, 4299, 6486, 6405, 4218, 4218, 4137, 1950, 3667, 1563, 73, 4514, 4441, 6455, 6373, 2106, 2031, 174, 2459, 4865, 4871, 4951, 4536, 4813, 1250, 1951, 557, 5093, 4444, 6454, 3456, 2112, 2160, 1208, 5988, 2263, 4606, 4531, 5420, 1775, 2018, 2099, 4286, 2021, 2140, 4299, 4299, 4299, 4299, 4299, 4299, 6486, 6486, 6486, 6486, 6486, 6405, 4218, 6162]

Length: **59** Speech Tokens

CoSTA-EN

Text Content: The sky is blue.

Speech Tokens:

[3704, 9189, 6657, 1764, 2990, 4447, 7478, 3456, 7246, 4514, 5348, 10357, 5269, 4813, 7115, 4858, 4449, 6390, 1776, 8886, 1450, 6476, 5635, 4771, 4772, 2585, 398, 11408, 2096, 6734, 6672, 43408]

Expanded Speech Tokens:

[3704, 4299, 4299, 4218, 4218, 4218, 4299, 4299, 1764, 2990, 4447, 6455, 6380, 3456, 2112, 84, 4514, 5348, 6320, 6158, 5269, 4813, 1277, 2006, 4858, 4449, 6390, 1776, 2148, 2166, 1450, 6476, 5635, 4771, 4772, 2585, 398, 1370, 2102, 2096, 4299, 4299, 4299, 4299, 4299, 4299, 6486, 6486, 6486, 4299, 4218, 4218, 4218, 4218, 4218, 4218, 4218, 4218, 4299, 6078]

Length: **32** BPE Speech Tokens / **60** Expanded Speech Tokens

Compression Rate: **53.33%**

Case Study 3: Sentence Rewriting

Input: *The order in which you add vectors does not matter. Please rewrite in simple and same meaning.*

Step-Audio 2 mini

Text Content: The order of adding vectors doesn't matter.

Speech Tokens:

[1499, 4218, 6486, 4299, 2112, 57, 2510, 566, 1763, 2032, 1358, 2906, 5102, 5815, 6058, 1518, 6067, 6059, 5329, 6050, 1760, 1001, 4970, 4275, 6453, 1517, 1951, 3650, 5834, 3647, 5837, 138, 4445, 5047, 6181, 4235, 4284, 6534, 110, 5918, 4570, 1788, 1869, 2247, 3105, 3098, 1685, 1688, 1697, 5912, 4671, 6374, 6373, 4245, 4299, 2085, 2253, 2, 2702, 5653, 5652, 4438, 5853, 1744, 1803, 4068, 5105, 5194, 5508, 5185, 867, 3146, 1676, 1679, 3872, 3847, 3729, 6486, 6486, 6405, 3975]

Length: **81** Speech Tokens

Vanilla-FT-EN

Text Content: When you add vectors, it doesn't matter which one comes first.

Speech Tokens:

[3704, 4299, 6486, 6405, 4218, 4218, 4218, 2121, 4316, 4346, 3644, 2567, 4379, 4471, 4966, 4885, 3482, 4130, 6071, 1517, 1978, 764, 4463, 4382, 5111, 4459, 57, 2417, 2152, 2074, 2097, 2219, 5191, 4536, 2112, 162, 1113, 4431, 3012, 3116, 1676, 1679, 1688, 1616, 5911, 4672, 6134, 762, 4056, 3975, 4299, 4299, 1124, 5102, 4697, 1536, 2049, 1770, 59, 3650, 4456, 5412, 5418, 4494, 5937, 2733, 132, 4074, 1639, 2195, 5108, 4540, 138, 3824, 1658, 1685, 1928, 1454, 2186, 2185, 2912, 2408, 4849, 438, 4758, 3864, 1380, 4344, 5101, 5029, 4862, 11, 50, 1726, 1805, 1833, 165, 5346, 5356, 5591, 666, 6498, 5409, 6374, 2916, 6537, 6534, 5780, 5320, 3863, 1685, 1619, 4509, 6376, 2139, 2247, 2193, 736, 4299, 4299, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 4299, 4218, 4218, 4218, 4218, 4218, 4218, 4218, 4299, 6159]

Length: **138** Speech Tokens

CoSTA-EN

Text Content: The order doesn't matter when you add vectors.

Speech Tokens:

[3704, 9189, 6676, 1725, 14093, 137, 8715, 2789, 2708, 4976, 4842, 297, 4601, 6864, 13239, 5189, 5348, 14410, 4439, 13344, 27059, 11544, 189, 2332, 7868, 2171, 18623, 110, 2276, 4643, 3454, 4129, 4367, 9321, 5435, 5920, 5108, 57, 1855, 2073, 1854, 73052, 11434, 7732, 3078, 3116, 11361, 3725, 4683, 4673, 7404, 1731, 6678, 20923, 6078]

Expanded Speech Tokens:

[3704, 4299, 4299, 4218, 4218, 4218, 6405, 6405, 1725, 1941, 63, 137, 2003, 2006, 2789, 2708, 4976, 4842, 297, 4601, 6059, 6058, 1689, 57, 5189, 5348, 5409, 5652, 4439, 5853, 5857, 4068, 173, 5192, 5917, 189, 2332, 3863, 1685, 2171, 1456, 1456, 110, 2276, 4643, 3454, 4129, 4367, 1763, 1979, 5435, 5920, 5108, 57, 1855, 2073, 1854, 2195, 5998, 2383, 6486, 2112, 4431, 3078, 3116, 1685, 1697, 3725, 4683, 4673, 5644, 5643, 1731, 4299, 4299, 4299, 4299, 4299, 4299, 4299, 4218, 4218, 4218, 4218, 4218, 4218, 4218, 4218, 4218, 4218, 4218, 4218, 6078]

Length: **55** BPE Speech Tokens / **94** Expanded Speech Tokens

Compression Rate: **58.51%**

Case Study 4: Open-ended Dialogue

Input: *What is the effect of saying hello to the person?*

Step-Audio 2 mini

Text Content: Saying "Hello" is an act of greeting, which can have several effects on both parties involved.

Speech Tokens:

[1499, 4218, 6486, 4299, 3894, 5158, 6455, 6458, 3222, 2246, 2195, 2573, 2825, 2834, 2753, 2807, 3995, 6182, 6181, 3984, 2031, 5835, 5838, 3651, 2112, 5109, 5832, 2918, 1046, 3962, 6148, 2486, 4780, 4806, 5077, 2813, 635, 728, 1457, 3644, 5827, 2028, 3645, 2922, 3165, 2032, 1951, 557, 4526, 6382, 3231, 2243, 3666, 1480, 1735, 1708, 5834, 5915, 5914, 4543, 1869, 165, 1896, 4431, 1520, 1978, 515, 5078, 4284, 6534, 2109, 1221, 657, 4367, 6071, 5587, 5100, 231, 75, 2816, 620, 1808, 4076, 3994, 4227, 4299, 3975, 5919, 5916, 3648, 3888, 4131, 4215, 4299, 2113, 3643, 5582, 240, 5244, 3135, 1788, 2355, 4390, 19, 1479, 5106, 5841, 3647, 326, 2169, 4509, 6374, 6377, 5652, 4406, 4537, 4356, 6543, 6310, 5086, 5097, 2003, 1028, 4519, 6453, 6537, 6540, 2656, 4540, 2924, 737, 197, 1203, 813, 4191, 6131, 5402, 5644, 1759, 1979, 5591, 4874, 587, 4314, 2130, 2139, 712, 626, 5087, 5094, 4770, 6375, 1893, 2166, 5670, 5589, 5753, 6266, 6050, 957, 2267, 566, 1295, 2996, 6373, 795, 2006, 4769, 1638, 4073, 2074, 6546, 713, 4946, 5022, 5053, 2789, 629, 719, 2915, 4368, 6507, 1032, 2241, 732, 4299]

Length: **186** Speech Tokens

Vanilla-FT-EN

Text Content: Saying hello greets them warmly, fostering connection and mutual respect. A simple gesture that often softens interactions, especially with those we've not met before.

Speech Tokens:

[3704, 6486, 6486, 4299, 4218, 5133, 5401, 6377, 5648, 4451, 2222, 4841, 5102, 4994, 5047, 6180, 4642, 5106, 5832, 5842, 3647, 3719, 5986, 47, 4457, 4807, 4997, 623, 632, 1370, 2184, 489, 1450, 4367, 3872, 4598, 4857, 5100, 1194, 4679, 4435, 1734, 1833, 1035, 59, 2216, 2124, 4316, 4319, 2130, 1430, 1457, 2870, 6482, 6238, 4311, 6174, 3206, 4043, 29, 386, 647, 1376, 2807, 2517, 6486, 6486, 3975, 5838, 5835, 3645, 2112, 6426, 6534, 3537, 491, 4942, 4411, 5644, 2139, 4431, 2290, 6059, 6554, 6302, 5336, 5075, 6181, 6502, 1815, 168, 2203, 1479, 3669, 2228, 5920, 4539, 1869, 3108, 5319, 4593, 159, 143, 32, 1482, 1484, 1483, 1735, 1951, 2924, 4382, 750, 1725, 161, 2127, 2128, 1646, 3563, 5342, 4851, 951, 6055, 6051, 3099, 4772, 5009, 5000, 5086, 4284, 4366, 3863, 4499, 4428, 6376, 1896, 2166, 4511, 5111, 1466, 815, 1923, 921, 3054, 1275, 4432, 2193, 735, 6486, 6486, 6486, 6486, 6486, 6486, 6405, 6162, 3894, 3648, 1950, 1703, 2918, 2254, 6455, 6380, 807, 4445, 6012, 2175, 3591, 5065, 5005, 885, 2490, 3147, 2330, 4459, 4414, 4672, 2814, 5487, 6054, 1673, 1679, 1676, 1685, 3875, 3813, 3651, 5919, 2922, 1545, 2085, 63, 2195, 220, 1516, 2003, 3407, 4949, 5026, 3429, 6507, 1032, 2217, 4495, 5124, 6099, 5645, 6374, 5643, 4727, 4945, 5053, 6345, 6537, 6453, 146, 68, 1480, 1481, 3671, 6385, 5404, 1763, 1285, 5021, 3751, 1887, 4515, 1632, 4115, 1598, 1493, 5921, 5998, 2379, 1569, 6021, 5322, 3138, 890, 65, 1477, 3668, 5858, 5402, 5405, 33, 2031, 5109, 3732, 5838, 5838, 1464, 1978, 1490, 4379, 4441, 6373, 2112, 2079, 2189, 5188, 4498, 4590, 4509, 4654, 4910, 4934, 4442, 5020, 4285, 1455, 1208, 2411, 4444, 6130, 1959, 2100, 73, 2300, 4835, 2810, 641, 2099, 1856, 5909, 5653, 2970, 1788, 5916, 2919, 1701, 3570, 3643, 3314, 4859, 3067, 6264, 1722, 3670, 1476, 5105, 5671, 5022, 111, 2733, 2220, 1641, 4071, 911, 3653, 5108, 138, 4920, 924, 2163, 3157, 5750, 2986, 6507, 6534, 2884, 5084, 2906, 719, 680, 1865, 1649, 3836, 3844, 3726, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 4299, 4299, 4299, 6159]

Length: **350** Speech Tokens

CoSTA-EN

Text Content: Saying hello fosters connection and breaks ice, especially meaningful from experience.

Speech Tokens:

[2975, 9039, 1950, 5136, 5404, 7084, 2982, 2267, 2573, 4742, 9727, 4885, 7184, 5839, 5, 3722, 5989, 47, 4540, 4726, 2729, 16524, 728, 725, 2028, 6768, 6991, 8720, 685, 10458, 13823, 7472, 2196, 3107, 11971, 4531, 46605, 7046, 93, 747, 5856, 2233, 9501, 1896, 141, 9701, 10551, 6955, 3668, 1501, 6776, 6949, 11175, 3671, 2155, 2175, 10911, 7151, 9950, 2904, 7161, 6912, 801, 1735, 1948, 7536, 4871, 731, 12873, 3563, 4510, 5405, 10063, 762, 6710, 8056, 7153, 2582, 4535, 7282, 6921, 2297, 13982, 4404, 8154, 4682, 4910, 5744, 2252, 5020, 913, 15772, 944, 5021, 4821, 6015, 4480, 4966, 4238, 2076, 6674, 716, 8929, 2915, 4278, 8502, 3872, 3242, 26855, 3586, 4312, 7604, 4849, 1140, 24474, 1839, 11667, 5101, 5335, 6293, 11772, 152, 14798, 8080, 5421, 8485, 5407, 771, 6669, 28091]

Expanded Speech Tokens:

[2975, 4299, 6486, 6486, 6486, 6486, 4299, 4218, 1950, 5136, 5404, 6376, 6377, 2982, 2267, 2573, 4742, 5048, 5047, 4885, 5352, 5109, 5839, 5, 3722, 5989, 47, 4540, 4726, 2729, 551, 635, 728, 725, 2028, 3888, 3888, 1701, 2112, 6426, 6534, 6534, 685, 4949, 5022, 4410, 6454, 2112, 2244, 2196, 3107, 3863, 6062, 4531, 5725, 3213, 2112, 87, 93, 747, 5856, 2233, 5921, 4539, 1896, 141, 5319, 4593, 2328, 890, 59, 1479, 1484, 3668, 1501, 2112, 1736, 4463, 4382, 49, 3669, 3671, 2155, 2175, 2179, 4367, 6302, 6050, 4598, 4849, 2904, 4299, 168, 5158, 6374, 801, 1735, 1948, 6320, 5671, 4871, 731, 359, 647, 3563, 4510, 5405, 6133, 6376, 762, 4299, 4299, 4299, 4218, 3651, 2922, 2112, 1951, 2582, 4535, 5158, 6376, 2139, 2166, 2297, 5191, 4543, 4404, 5319, 4509, 4682, 4910, 5744, 2252, 5020, 913, 2131, 2127, 944, 5021, 4821, 6015, 4480, 4966, 4238, 2076, 6534, 6534, 716, 626, 626, 2915, 4278, 6534, 6534, 4356, 3872, 3242, 5591, 4955, 3586, 4312, 1763, 1277, 4849, 1140, 3024, 6373, 1839, 2166, 954, 5101, 5335, 6293, 4358, 1685, 152, 386, 35, 1480, 3671, 5421, 5402, 5405, 5407, 771, 4299, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 6486, 4299, 4299, 4299, 4299, 4299, 4299, 4299, 4299, 4299, 4299, 6159]

Length: **130** BPE Speech Tokens / **207** Expanded Speech Tokens

Compression Rate: **62.80%**