# PriKE-ICL: Enhancing In-Context Learning Ability via Prior Knowledge Steering in Neural Codec Language Model

**Anonymous ACL submission**

## Abstract

The neural codec language model (VALL-E) demonstrates emergent in-context learning capabilities, achieving impressive naturalness and enabling zero-shot adaptation with just a 3-second speech prompt. However, in-context learning has demonstrated limited effectiveness across various settings, occasionally leading to speech with mispronunciations, word skipping, and repetitions. Additionally, it poses challenges in synthesizing high-quality personalized speech, particularly under conditions of reverberation and background noise in the acoustic prompt. To overcome these limitations, we propose a novel method called PriKE-ICL, a generative prior knowledge steering neural codec language model. PriKE-ICL integrates "diagonal", "disentangle" and "diverse" prior knowledge into the generation process while maintaining the decoder-only transformer architecture. First, we incorporate guided *diagonal* attention into the learning of text-speech alignment, explicitly indicating the monotonic generation process. Furthermore, an acoustic prompt disentanglement is proposed to *disentangle* the timbre prompt from the speech content, creating a smoother space of acoustic prompt representations. During inference time, we propose to retrieve acoustic examples that are phonemically or timbre-wise similar to a test query sample to formulate its corresponding prompt. Intuitively, examples selected with this strategy may offer more *diverse* inputs, unleashing the power of the Neural Codec Language Model in speech generation. PriKE-ICL surpasses the state-of-the-art zero-shot TTS model VALL-E in both intelligibility (5.80% vs. 3.82% word error rates) and audio similarity (0.562 vs. 0.638) in monolingual zero-shot TTS, can achieve multilingual and noise-conditioned zero-shot TTS without the need for additional training.

## 1 Introduction

Neural Large language models (LLMs) have shown significant advancements in the field of natural language generation (Brown et al., 2020; Radford et al., 2019). When the model size is sufficiently large, LLMs exhibit powerful in-context learning capabilities, allowing them to handle previously unseen tasks based on a text instruction (commonly referred to as a prompt) in a zero-shot or few-shot manner (Wei et al., 2022). Furthermore, the straightforward yet efficient next-token prediction task of LLMs facilitates their application in various domains, including vision (Dehghani et al., 2023) and speech synthesis (Wang et al., 2023a; Zhang et al., 2023; Xin et al., 2024; Han et al., 2024; Song et al., 2024; Du et al., 2024), provided that the data is transformed into discrete speech tokens. This paper centers on the language modeling component of text-to-speech (TTS) synthesis. Recent studies (Wang et al., 2023a; Zhang et al., 2023; Du et al., 2024) have demonstrated that text-to-speech (TTS) can be effectively modeled using a decoder-only language model by employing a neural codec (Défossez et al., 2024; Zeghidour et al., 2021) to transform continuous waveforms into discrete tokens. These methods are trained on extensive corpora consisting of thousands of hours of speech data (Kahn et al., 2020), covering over 7000 unique speakers. It has been shown that as the number of model parameters increases, in-context learning (ICL) becomes a notable feature of LLM-based TTS (Wang et al., 2023a; Xin et al., 2024). Performance in tasks such as zero-shot text-to-speech improves in terms of audio quality for previously unheard speakers, and larger models generate more natural and smooth speech. Unlike traditional learning methods that require updating model parameters, ICL enables good performance with a prompt that includes only a 3-second acoustic sample.

Despite the remarkable ICL ability of LLM-based TTS, the efficacy of ICL is uneven and may be highly sensitive to the choices of text and acoustic prompts (Xin et al., 2024; Du et al., 2024). This

creates barriers to developing LLM-based TTS applications that achieve both adaptability and robustness. Even the state-of-the-art model may encounter challenges: 1) Current methods (Wang et al., 2023a; Du et al., 2024) typically concatenate phoneme tokens and acoustic tokens into a single sequence to train language models. This approach relies entirely on the transformer's self-attention mechanism to learn the alignment between audio and phoneme sequences. However, this leads to instability since self-attention does not explicitly ensure a monotonic alignment between the audio and phoneme sequences. This results in certain words being unclear, missed, or duplicated in speech synthesis, especially for those words that occur infrequently in the training data. 2) The decoder-only language model architecture may encounter potential attention degradation issues (Fu et al., 2023). As the generated sequence lengthens, the alignment quality between the target audio sequence and the source phoneme sequence deteriorates, leading to inaccurate or lower-quality speech outputs. Further, in cases where the acoustic prompt speech contains reverberation or background noise, this leads to the leakage of information related to poorer audio quality content into the timbre, resulting in attention collapse (unintelligible gibberish when the model fails to focus on a single input token). 3) Autoregressive (AR) models generate the current tokens based solely on previous tokens without allowing user control during the inference process, which sometimes leads to semantic repetitions or incoherence in the generated output (Du et al., 2024). In the context of TTS tasks, models like VALL-E (Wang et al., 2023a) face similar challenges: they cannot precisely map which segment of the output audio corresponds to each prompt phoneme. As a result, there is no straightforward method to promptly detect and prevent issues during the generation process. These limitations manifest in various ways, such as meaningless phoneme repetitions, transpositions, omissions, or even catastrophic failures like infinite silence—where the model erroneously outputs silence or noise tokens for an extended period without stopping. These pose significant challenges to existing LLM-based TTS methods.

Although significant progress has been made, existing methods do not handle the challenges in VALL-E well. Specifically, 1) **multi-task data augmentation** methods (Wang et al., 2024, 2023c) enhance robustness by incorporating multitasking data, including noise suppression and noisy speech editing. However, these methods escalate training costs as multitasking data increases. Additionally, there is no dedicated strategy for enhancing in-context learning, and the synthesized speech continues to experience issues such as word skipping and repeating. 2) On the other hand, **prior knowledge guidance** methods construct shifting relative position embeddings for input phoneme sequences, introducing knowledge of relative positions to indicate the generation process (Du et al., 2024) or design multi-scale acoustic prompts to incorporate personal speaking style knowledge (Lei et al., 2024), aiming to enhance the naturalness and speaker similarity of synthesized speech. However, these references based on VALL-E fail to address scenarios where acoustically cued speech contains reverberation or background noise. Consequently, the quality of the synthesized speech degrades rapidly with the introduction of more noise.

To address the above challenges of VALL-E based methods, we propose a novel method called PriKE-ICL, a generative prior knowledge steering neural codec language model, which incorporates "diagonal", "disentangle" and "diverse" prior knowledge into the generation process while retaining the architecture of a decoder-only Transformer. We conduct experiments on the LibriTTS (Zen et al., 2019), MLS (Pratap et al., 2020) and noisy LibriTTS (Valentini-Botinhao et al., 2017) datasets. The results of extensive quantitative and qualitative evaluations show that in terms of speech quality and pronunciation accuracy, PriKE-ICL is on par with state-of-the-art VALL-E (Wang et al., 2023a; Du et al., 2024) based methods. Our main contributions are:

- We propose a novel generative prior knowledge steering neural codec language model architecture to enhance in-context learning. We design guided diagonal attention and acoustic prompt disentanglement to guide generation process.

- We propose two different scenarios of context-learning diversity enhancement methods, specifically tailored for zero-shot multilingual synthesis and zero-shot synthesis under noisy conditions, respectively. These methods unleash the power of the Neural Codec Language Model in speech generation.

- Extensive experimental results demonstrate the proposed method performs well against

state-of-the-art models on three benchmark datasets.

## 2 The Method

Figure. 1 illustrates an overview of the PriKE-ICL architecture. Building upon the principles introduced in VALL-E (Wang et al., 2023a), PriKE-ICL employs the "diagonal", "disentangle" and "diverse" prior knowledge into the neural codec language model. The model learns to perform conditional generation of a neural code sequence, denoted as $\mathcal{O}$, based on two input prompts: textual prompt $\mathcal{T}$ and acoustic prompt $\mathcal{S}$. The neural codes may also be referred to as acoustic tokens.

The textual prompt $\mathcal{T}$ is a sequence of phonemes obtained by applying grapheme-to-phoneme conversion to an input text. The textual prompt conveys the semantic information, and thus it is called semantic tokens. Conversely, the acoustic prompt $\mathcal{S}$ encapsulates the acoustic information of an input speech signal. It is obtained by converting the input audio into a sequence of acoustic tokens with an encoder of the neural codec model. The alignment between $\mathcal{T}$ and $\mathcal{O}$ denoted as $\mathcal{A}$, we introduce a rough assumption, which prompts the alignment $\mathcal{A}$ to be "diagonal" in Section 2.1. In addition, we think the $\mathcal{S}$ should "disentangle" from the $\mathcal{O}$, especially if the input audio has reverberation or an undercurrent of noise, we will explain it in Section 2.2. Furthermore, we incorporate "diverse" in the acoustic prompt by the idea of automatic chain-of-thought. The details will be explained in Section 2.3. The output $\mathcal{O}$ is a sequence of neural codes of the desired signal, which is then translated into a waveform signal with the codec decoder. We use EnCodec (Défossez et al., 2024) as the neural codec model, following the prior work (Wang et al., 2023a).

### 2.1 Diagonal

In general, an attention module is quite costly to train. Therefore, if there is some prior knowledge, incorporating them into the model may be a help to alleviate the heavy training. We show that the following simple method helps to train the attention module (Tachibana and Shunsuke, 2018) .

In TTS, the alignment matrix $\mathcal{A}$ lies in the very small subspace of $R^{M\times N}$. This is because of the rough correspondence of the order of the characters and the audio segments. That is, when one reads a text, it is natural to assume that the text position $m$ progresses nearly linearly to the time $n$, i.e., $m \sim an$, where $a \sim M/N$. This is a noticeable difference of TTS from other seq2seq learning techniques such as machine translation, where an attention module needs to resolve the word alignment between two languages that have very different syntax, e.g. English and Chinese.

Based on this idea, we propose the prior "diagonal" knowledge enhancing in-context learning, which aims to guide the alignment matrix $\mathcal{A}$ towards being "nearly diagonal":

$$\begin{aligned}\mathcal{L}_{AR} &= -\sum_{t=1}^{T} logP(o_{t,1}|\mathcal{T},\mathcal{S},o_{<t,1};\theta_{AR}) \\ &+E[\mathcal{A}(1-exp\{-(m/M-t/T)^2/2g^2\})], \\ \mathcal{L}_{NAR} &= -\sum_{l=2}^{8} logP(o_{:,l}|\mathcal{T},\mathcal{S},o_{:,<l};\theta_{NAR}) \\ &+E[\mathcal{A}(1-exp\{-(m/M-t/T)^2/2g^2\})],\end{aligned} \tag{1}$$

where $o_{<t,1} = [o_{1,1},\ldots,o_{t-1,1}]$, while $\theta_{AR}$ represents the AR Transformer model parameters. $\theta_{NAR}$ represents the NAR model parameters, while $o_{:,l}$ denotes the entire sequence of $o_{t,l}$ for the $l$th layer, and $o_{:,<l} = [o_{:,1},\ldots,o_{:,l-1}]$ for $l = 2,\ldots,8$. Note that the AR model in PriKE-ICL is conditioned on the concatenated embeddings of both the acoustic and textual prompts. This formulation differs from that of VALL-E, where the AR model is only conditioned on the textual prompt and the past acoustic history. In addition, in order for the single NAR model to process each of the seven layers, the acoustic tokens from the first to $(l-1)$th layers, $o_{:,<l}$, are embedded and summed up.

### 2.2 Disentangle

In VALL-E, the acoustic prompt $\mathcal{S}$ is the first layer acoustic token of the enrolled speech $\tilde{c}_{:,1}$. The acoustic prompt $\mathcal{S}$ and the output $\mathcal{O}$ is the same speech during the training, however, VALL-E is limited by the reverberation and background noise of the acoustic prompt, which leads to the potential leakage of information related to poorer audio quality content into the timbre. Based on this idea, we propose the prior "disentangle" knowledge enhancing in-context learning, which aims to disentangle the acoustic prompt $\mathcal{S}$ from the output $\mathcal{O}$.

For the disentanglement between $\mathcal{S}$ and $\mathcal{O}$. The variational contrastive log-ratio upper bound (vCLUB) (Cheng et al., 2020) is used to compute the upper bound of mutual information (MI) for irrelevant information of the speaker, decreasing

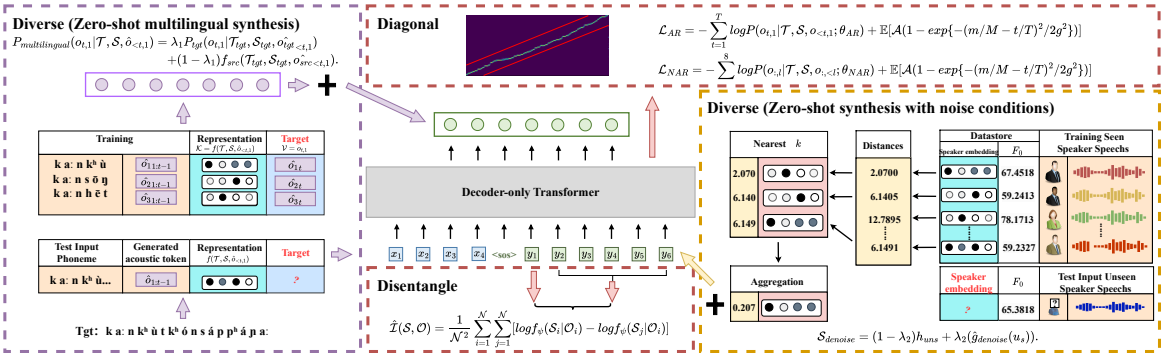


Figure 1: Framework of proposed model.

the correlation among different speaker-irrelevant speech representations:

$$\hat{\mathcal{I}}(\mathcal{S},\mathcal{O}) = \frac{1}{\mathcal{N}^2}\sum_{i=1}^{\mathcal{N}}\sum_{j=1}^{\mathcal{N}}[\log f_\psi(\mathcal{S}_i|\mathcal{O}_i) - \log f_\psi(\mathcal{S}_j|\mathcal{O}_i)], \quad (2)$$

where $\mathcal{N}$ represents the samples from $\mathcal{S}$ and $\mathcal{O}$. $f_\psi(\mathcal{S}|\mathcal{O})$ is a variational distribution with parameter $\psi$ to approximate $f(\mathcal{S}|\mathcal{O})$. $\hat{\mathcal{I}}$ is the unbiased estimator for vCLUB with samples $\{\mathcal{S}_i, \mathcal{O}_i\}$. The indexes $i$ and $j$ are the samples of $\mathcal{S}$ and $\mathcal{O}$. By minimizing Eq. 2, we can decrease the correlation among accent features $\mathcal{S}$ and speech features $\mathcal{O}$. The MI loss is:

$$\mathcal{L}_{MI} = \hat{\mathcal{I}}(\mathcal{S},\mathcal{O}), \quad (3)$$

in PriKE-ICL training stage, we add the MI loss to the $\mathcal{L}_{AR}$ and $\mathcal{L}_{NAR}$.

### 2.3 Diverse

In-context learning (ICL) enables neural codec language model to perform a target task by feeding a few (textual/acoustic) prompted examples as part of the input (Liu et al., 2023). Without gradient update, ICL allows a single model to perform various tasks universally. Despite the remarkable ICL ability of neural codec language model, the efficacy of ICL is uneven and can be highly sensitive to the choices of textual, acoustic prompt. However, in the domain of large language models (LLMs), several research avenues aim to enhance the performance of ICL: (i) Retrieving related demonstrations for the test instance, where the prevalent practice involves dynamically retrieving related training examples for a given test input. (ii) Augmenting with fine-grained information, such as incorporating task instructions. (iii) Manipulating output probabilities of LLMs rather than directly computing the likelihood of target labels. Inspired by research on enhancing LLMs' in-context learning, sampling diverse prompts may help mitigate the impact of misleading information, even though a small portion (e.g., 1/8) of errors may still exist in the prompts (Liu et al., 2022). We propose two different scenarios of context-learning diversity enhancement methods suitable for zero-shot multilingual synthesis and zero-shot synthesis with noise conditions, respectively.

**Zero-shot multilingual synthesis.** In zero-shot multi-lingual speech synthesis, the goal is to synthesize the speech of another language without updating model parameters, which is more challenging than conventional monolingual TTS. We design a $k$-nearest-neighbor zero-shot multilingual in-context learning method ($k$NN-mICL). $k$NN-mICL involves augmenting the output probabilities of a pre-trained monolingual neural codec language model with a nearest neighbor retrieval mechanism, allowing the model direct access to a datastore of cached examples. The synthesizer is generated token-by-token; at each time step, we use the multilingual G2P and find the most similar textual prompt in the datastore, and compute a distribution over the corresponding target tokens. This distribution is then interpolated with the output distribution from the pre-trained neural codec language model.

More specifically, given inputs textual prompt $\mathcal{T}$ and acoustic prompt $\mathcal{S}$, a AR neural codec language model outputs a sequence of tokens $o_{AR} = [o_{1,1}, \ldots, o_{t-1,1}, o_{t,1}]$, a NAR neural codec language model outputs a sequence of tokens $o_{NAR} = [o_{:,1}, \ldots, o_{:,l-1}, o_{:,l}]$. When using autoregressive decoders, the output distribution for each acoustic token $o_{t,1}$ is conditioned on the entire textual and

acoustic prompts sequence as well as the previous acoustic tokens, $P(o_{t,1}|\mathcal{T}, \mathcal{S}, o_{<t,1})$. The datastore consists of a set of key-value pairs and is defined as follows:

$$(\mathcal{K}, \mathcal{V}) = \{f(\mathcal{T}, \mathcal{S}, o_{<t,1}), o_{t,1}\}, \quad (4)$$

where $f$ represents a mapping from input to an intermediate representation of the monolingual pre-trained neural codec language model. At test time, given a target language inputs textual prompt $\mathcal{T}_{tgt}$ and acoustic prompt $\mathcal{S}_{tgt}$, the model outputs a distribution $P_{tgt}(o_{t,1}|\mathcal{T}_{tgt}, \mathcal{S}_{tgt}, \hat{o}_{<t,1})$ for the acoustic token $o_{t,1}$ at every step of generation, where $\hat{o}$ represents the generated tokens. The model also outputs the representation $f_{src}(\mathcal{T}_{tgt}, \mathcal{S}_{tgt}, \hat{o_{src}}_{<t,1})$, which is used to query the datastore for the $k$ nearest neighbors $N$ according to squared-$L^2$ distance, $d$. The model and $kNN$ distributions are interpolated with a tuned parameter $\lambda_1$, resulting in the final $k$NN-mICL distribution:

$$\begin{aligned} &P_{multilingual}(o_{t,1}|\mathcal{T}, \mathcal{S}, \hat{o}_{<t,1}) \\ &= \lambda_1 P_{tgt}(o_{t,1}|\mathcal{T}_{tgt}, \mathcal{S}_{tgt}, \hat{o_{tgt}}_{<t,1}) \\ &+(1-\lambda_1) f_{src}(\mathcal{T}_{tgt}, \mathcal{S}_{tgt}, \hat{o_{src}}_{<t,1}). \end{aligned} \quad (5)$$

**Zero-shot synthesis with noise conditions.** First, we employ the pre-trained speaker verification model (Wang et al., 2023b) as a speaker encoder, this choice allows us to harness its robust speaker authentication capabilities, even in the presence of noisy audio conditions. Then we utilize a substantial datastore of audio samples from seen speakers by extracting both the fundamental frequency (F0) features and the corresponding speaker embeddings. Next the non-parametric speaker adaptor generates the unseen speaker embedding based on the datastore using a nearest neighbour retrieval mechanism, and finally the unseen speaker embedding is combined with text and acoustic prompt embeddings. Given an input speech of seen speaker $u_s$. Our datastore is established offline and comprises a collection of key-value pairs:

$$(\mathcal{K}, \mathcal{V}) = \{f(u_{si}), g(u_{si}), \forall u_{si} \in u_s\}, \quad (6)$$

where $f_{denoise}$ represents a mapping from input speech to a discrete F0 representation of the speaker encoder, $g_{denoise}$ represents a mapping from input speech to a high-dimensional speaker representation.

At test time, given a target unseen speaker $u_{uns}$, the pre-trained speaker encoder model outputs the speaker embedding distribution $h_{uns}$ for the target unseen speaker $u_{uns}$. The model also outputs the F0 represention $P_{uns}$, which is used to query the datastore for the $k$ nearest neighbors $\mathcal{N}$ according to squared-$L^2$ distance, $d$. We have:

$$\mathcal{S}_{denoise} = (1-\lambda_2)h_{uns} + \lambda_2(\hat{g}_{denoise}(u_s)). \quad (7)$$

## 3 Metrics

We continue the evaluation in the manner of (Wang et al., 2023a), focusing on Correctness and Intelligibility, Coherence, and Diversity and Quality aspects. For Correctness and Intelligibility, we assess the synthesized speech's word error rate (WER) compared to the input text, utilizing public automatic speech recognition (ASR) models for comparability. In English-only setups, we employ the HuBERT-L (Hsu et al., 2021) model pre-trained on 60K hours of Librilight (Kahn et al., 2020) and fine-tuned on 960 hours of Librispeech (Panayotov et al., 2015). For multilingual setups, we utilize the Whisper (Cao et al., 2012) large-v2 model.

Regarding Coherence, we utilize the WavLM-TDCNN (Chen et al., 2022) speaker embedding model to measure the similarity between the embedding of the generated speech and that of the audio context. Different embedding models may reflect coherence across various attributes. We report the similarity with respect to the resynthesized audio context by its vocoder (SIM-r) and the similarity against the original audio context (SIM-o).

Additionally, we utilize Mean Opinion Score (MOS) for subjective audio quality evaluation (QMOS), and Similarity MOS (SMOS) for subjective audio similarity evaluation, where pairs of prompt and system-generated audio clips are rated on a scale of 1 to 5, with 5 representing the highest quality or similarity.

## 4 Experiment

**Data:** For monolingual models, we conduct experiments using the LibriTTS (Zen et al., 2019) dataset , which is a multi-speaker transcribed English speech dataset. Its training set comprises approximately 580 hours of speech data from 2,306 speakers. For zero-shot multilingual synthesis tasks, we use the MLS test (Pratap et al., 2020) and we employ the epitran multi-language G2P tool. We also validate our approach using 100 audio samples from the Aishell-3 (Shi et al., 2021)

| Methods | **WER** | **SIM-o** | **SIM-r** | **QMOS** | **SMOS** |
|---|---|---|---|---|---|
| Ground Truth | 1.92 | 0.698 | n/a | 3.89($\pm$0.18) | 3.92($\pm$0.16) |
| Encodec resynthesis | 2.08 | 0.671 | n/a | 3.80($\pm$0.12) | 3.87($\pm$0.11) |
| *cross-sentence* | | | | | |
| VALL-E | 5.80 | - | 0.562 | - | - |
| YourTTS | 7.94 | 0.283 | n/a | 3.21($\pm$0.11) | 3.09($\pm$0.15) |
| VALL-T | 4.16 | 0.608 | 0.613 | 3.69($\pm$0.13) | 3.60($\pm$0.10) |
| **PriKE-ICL** | **3.82** | **0.625** | **0.638** | **3.74**($\pm$0.12) | **3.69**($\pm$0.13) |
| *continuation* | | | | | |
| VALL-E | 5.01 | 0.407 | 0.421 | - | - |
| VALL-T | 4.05 | 0.446 | 0.461 | - | - |
| **PriKE-ICL** | **3.71** | **0.542** | **0.571** | - | - |

Table 1: Comparison of WER, SIM-o, SIM-r, QMOS and SMOS of the proposed PriKE-ICL and the previous works on the LibriTTS datasets.

| | **Ref** | Fr | | En | | Vi | | Lao | | My | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **WER** | **SIM-o** | **WER** | **SIM-o** | **WER** | **SIM-o** | **WER** | **SIM-o** | **WER** | **SIM-o** |
| GT | - | 6.9 | 0.698 | 6.1 | 0.701 | 9.6 | 0.612 | 10.3 | 0.590 | 12.6 | 0.521 |
| YT | Fr | 15.2 | 0.257 | 8.1 | 0.323 | n/a | n/a | n/a | n/a | n/a | n/a |
| | En | 16.1 | 0.221 | 7.8 | 0.370 | n/a | n/a | n/a | n/a | n/a | n/a |
| | Vi | 16.8 | 0.219 | 8.2 | 0.319 | n/a | n/a | n/a | n/a | n/a | n/a |
| | Lao | 16.2 | 0.220 | 8.5 | 0.298 | n/a | n/a | n/a | n/a | n/a | n/a |
| | My | 16.6 | 0.211 | 8.9 | 0.211 | n/a | n/a | n/a | n/a | n/a | n/a |
| | **AVG** | 16.2 | 0.226 | 8.3 | 0.304 | n/a | n/a | n/a | n/a | n/a | n/a |
| **PriKE-ICL** | Fr | 8.8 | 0.528 | 7.1 | 0.519 | 11.6 | 0.439 | 15.3 | 0.481 | 15.5 | 0.446 |
| | En | 8.9 | 0.401 | 7.4 | 0.503 | 12.1 | 0.421 | 16.8 | 0.436 | 18.3 | 0.386 |
| | Vi | 9.1 | 0.459 | 7.3 | 0.467 | 13.7 | 0.420 | 15.4 | 0.461 | 15.2 | 0.413 |
| | Lao | 9.0 | 0.409 | 7.9 | 0.481 | 13.2 | 0.459 | 17.2 | 0.470 | 17.1 | 0.415 |
| | My | 9.3 | 0.480 | 7.8 | 0.501 | 14.2 | 0.469 | 17.8 | 0.441 | 18.7 | 0.409 |
| | **AVG** | **9.0** | **0.455** | **7.5** | **0.494** | **13.0** | **0.442** | **16.5** | **0.458** | **17.0** | **0.414** |

Table 2: Multilingual zero-shot TTS results on filtered MLS test sets. GT/YT/PriKE-ICL refers to ground truth/YourTTS/multilingual PriKE-ICL. "Ref" column shows the audio context language.

test set. Additionally, we conduct experiments with 100 audio samples from languages with significant morphological differences, such as Vietnamese and Laotian. For zero-shot synthesis with noise conditions tasks, we utilize open-source datasets containing noise (Valentini-Botinhao et al., 2017). We employ the Wespeaker (Wang et al., 2023b) voice recognition model.

**Model:** We utilize our Encodec (Défossez et al., 2024) speech tokenizer with a frame shift of 20ms and an output waveform sampling rate of 16k, consisting of 8 residual vector quantization (RVQ) indices for each frame. To ensure a fair comparison between VALL-E and our proposed model PriKE-ICL, we adopt the approach introduced in PriKE-ICL. This involves predicting the sequence of the first RVQ index with auto-regressive models and then predicting the remaining 7 RVQ indices conditioned on the first RVQ index with a separate non-auto-regressive (NAR) model. PriKE-ICL shares an identical architecture with VALL-E, featuring 12 layers of Transformer blocks, each comprising 12 attention heads and a hidden dimension of 1024.

**Training:** We train our model for 40 epochs using a ScaledAdam optimizer (Yao et al., 2024). The learning rate scheduler employed is Eden, with a base learning rate of 0.05, an epoch scheduling factor of 4, and a step scheduling factor of 5000.

**Baselines:** We consider three baselines: 1) VALL-E (Wang et al., 2023a), SOTA for English zero-shot

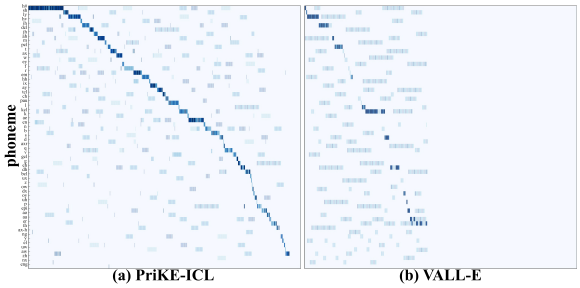


Figure 2: Alignment analysis.

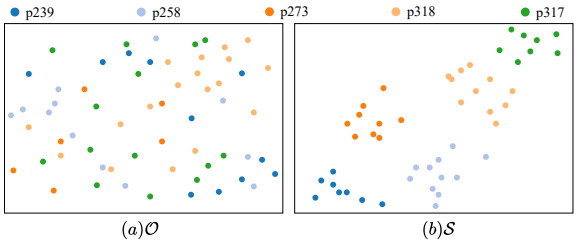


Figure 3: Acoustic prompt analysis.

TTS . 2) YourTTS (Casanova et al., 2022), SOTA multilingual zero-shot TTS model. 3) VALL-T (Du et al., 2024), the latest enhancement for VALL-E involves replacing the Transformer architecture with Transducer.

### 4.1 Alignment analysis

In our experiment, when we introduced the "diagonal" knowledge into the objective, Figure 3(b) showcases an audio sample with word repetition from VALL-E due to the absence of a "diagonal" relationship. In this instance, a phoneme segment is spoken twice, separated by a short silence. In contrast, when this utterance is processed by PriKE-ICL, the generation is correct without any repetition, as demonstrated in Figure 3(a).

### 4.2 Acoustic prompt analysis

Specifically, we randomly select five speakers from the VCTK (Veaux et al., 2016) test dataset and pick 10 random speech samples per speaker. We extract the acoustic prompt from the enrolled speech $\tilde{c}$. The first layer acoustic token of the $\tilde{c}_{:,1}$ is denoted as $\mathcal{S}$, while the all acoustic tokens of the $\tilde{c}$is denoted as $\mathcal{O}$. By performing mean pooling along the temporal dimension, each representation is converted into a single vector. These vectors are then visualized in a 2D space using t-SNE, with speech samples from the same speaker represented in the same color. From the visualization, it can be observed that the $\mathcal{O}$ representations for different speakers are scattered randomly without a discernible pattern. In contrast, the $\mathcal{S}$ representations for the same speaker tend to cluster together, while remaining distinct from those of other speakers. This suggests that the $\mathcal{S}$ representation space, guided by decoupled knowledge, exhibits a smoother distribution, which is more conducive to achieving zero-shot speech synthesis.

### 4.3 Monolingual and multilingual zero-shot TTS

Table 1 presents the zero-shot TTS outcomes for the English model PriKE-ICL. As per the approach outlined in (Wang et al., 2023a), the test set comprises samples ranging from 4 to 10 seconds in length, drawn from Librispeech test-clean. We investigate two prompting strategies: cross-sentence prompting, utilizing a 3-second clip from another sample of the same speaker as the audio context, and continuation prompting, employing the initial 3 seconds of each utterance. PriKE-ICL demonstrates superior performance across all metrics in both scenarios compared to all baselines. Particularly noteworthy is PriKE-ICL's significantly more effective style transfer (+0.101/+0.108 SIM-r on cross sentence/continuation) compared to VALL-E, with an even more pronounced performance gap observed when compared against raw audio (+0.141 SIM-o on continuation). MOS studies further validate the subjective superiority of PriKE-ICL in terms of quality and similarity over YourTTS.

Table 2 presents multi-lingual zero-shot TTS results, where the training and the inference datasets are in different languages. Note that PriKE-ICL is not trained on any sample with multiple languages in an utterance spoken by the same speaker. The test set is constructed using filtered MLS test split. For each target text, we sample one 3-second long audio context from each language, which creates 36 language transfer directions in total. PriKE-ICL yields better performance than YourTTS everywhere. Specifically, on En/Fr/Pt which YourTTS supports, PriKE-ICL obtains 3.1%/5.9%/8.1% lower WERs and 0.136/0.141/0.160 higher similarity averaged across audio context in nine languages.

### 4.4 Zero-shot synthesis with noise conditions

In the context of noisy environments, we artificially add noise to the LibriTTS test-clean dataset (Valentini-Botinhao et al., 2017). We use

| Methods | **WER** | **SIM-o** | **SIM-r** | **QMOS** | **SMOS** |
|---|---|---|---|---|---|
| Ground Truth | 1.92 | 0.698 | n/a | 3.89($\pm$0.18) | 3.92($\pm$0.16) |
| Encodec resynthesis | 2.08 | 0.671 | n/a | 3.80($\pm$0.12) | 3.87($\pm$0.11) |
| *reverberation* | | | | | |
| VALL-E | 11.34 | - | 0.412 | - | - |
| YourTTS | 15.21 | 0.210 | n/a | 3.02($\pm$0.12) | 2.87($\pm$0.10) |
| VALL-T | 9.51 | 0.556 | 0.581 | 3.35($\pm$0.14) | 2.92($\pm$0.11) |
| **PriKE-ICL** | **5.79** | **0.602** | **0.587** | **3.50**($\pm$0.13) | **3.41**($\pm$0.12) |
| *background noise* | | | | | |
| VALL-E | 10.29 | - | 0.539 | - | - |
| YourTTS | 13.19 | 0.297 | n/a | 3.21($\pm$0.11) | 3.01($\pm$0.12) |
| VALL-T | 7.23 | 0.583 | 0.601 | 3.51($\pm$0.10) | 3.45($\pm$0.12) |
| **PriKE-ICL** | **6.16** | **0.591** | **0.619** | **3.62**($\pm$0.11) | **3.57**($\pm$0.10) |

Table 3: Comparison of WER, SIM-o, SIM-r, QMOS and SMOS of the proposed model and the previous works on the noisy LibriTTS datasets.

the noisy speech as acoustic prompts input into different models to achieve zero-shot speech synthesis. We set up two different noise scenarios reflecting real-world situations: 1) Reverberation scenario: We play the original dataset audio files from a smartphone and capture the audio using another microphone. Finally, we save the captured speech to create audio files with natural reverberation noise. 2) Background noise scenario: We create a noisy test set by mixing the filtered Librispeech test-clean with non-speech noise, ensuring that it overlaps with 50% of the duration at a -10dB signal-to-noise ratio.

Table 3 presents the comparison between PriKE-ICL and other baseline models under these two noise scenarios. It can be observed that in the reverberation scenario, the performance of other baseline methods significantly declines, while our model achieves the best results. This indicates that the knn context learning enhancement method based on voiceprint recognition models can effectively handle cases where the acoustic prompt quality is poor. Additionally, under the condition of acoustic prompts containing background noise, our model also achieves the best performance.

### 4.5 Ablation on generative modeling approaches

We conduct an ablation study to analyze the performance effects of different methods for PriKE-ICL, with results presented in Table 4. From the findings, it is observed that when the "diagonal" prior knowledge steering is removed, the model still performs well and outperforms most of the baseline models, but there is a decrease in speech naturalness. Additionally, when the "disentangle" component is removed, there is a significant decrease in audio similarity. Moreover, without the "diverse" component, the results are poor, and the model no longer performs well in the zero-shot multilingual synthesis and zero-shot synthesis with noise conditions tasks.

| Methods | **WER** | **SIM-o** | **SIM-r** |
|---|---|---|---|
| Ground Truth | 1.92 | 0.698 | n/a |
| Encodec resynthesis | 2.08 | 0.671 | n/a |
| VALL-E | 5.80 | - | 0.562 |
| PriKE-ICL | 3.82 | 0.625 | 0.638 |
| PriKE-ICL *w/o* "diagonal" | 4.80 | 0.609 | 0.611 |
| PriKE-ICL *w/o* "disentangle" | 4.03 | 0.580 | 0.601 |
| PriKE-ICL *w/o* "diverse" | 5.12 | 0.545 | 0.570 |

Table 4: Comparison of WER, SIM-o, SIM-r, QMOS and SMOS of the ablation study. Each model is evaluated on the zero-shot TTS task.

## 5 Conclusion

This paper introduces PriKE-ICL, an interpretable augmented TTS model for contextual learning. It leverages three types of prior knowledge to address challenges in traditional VALL-E models. The "diagonal' prior improves speech-phoneme alignment, solving issues like missing words. The "disentangle" prior smooths speech representation for better generalization, and the "diverse" prior uses $k$NN to handle zero-shot multilingual synthesis and noisy conditions. PriKE-ICL matches state-of-the-art performance in zero-shot TTS and noisy synthesis without additional training.

## 6 Limitation

The primary limitation of the current work is its reduced performance in highly noisy and reverberant environments, where the model struggles to maintain speech quality and accent consistency. Additionally, while the model successfully disentangles accent and speaker features in most cases, it may face challenges when applied to accents or speakers not sufficiently represented in the training data, limiting its generalizability in such scenarios. Further improvements are needed to enhance robustness across a wider range of acoustic conditions and speaker-accent combinations.

## References


Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the NeurIPS Conference 2020*, pages 1877–1901.

Nan Cao, Yu-Ru Lin, Xiaohua Sun, David Lazer, Shixia Liu, and Huamin Qu. 2012. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *IEEE transactions on visualization and computer graphics*, 18(12):2649–2658.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *Proceedings of the ICML Conference 2022*, pages 2709–2720.

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

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *Proceedings of the ICML Conference 2020*, pages 1779–1788.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2024. High fidelity neural audio compression. In *Proceedings of the ICLR Conference 2024*, pages 2835–8856.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. 2023. Scaling vision transformers to 22 billion parameters. In *Proceedings of the ICML Conference 2023*, pages 7480–7512.

Chenpeng Du, Yiwei Guo, Hankun Wang, Yifan Yang, Zhikang Niu, Shuai Wang, Hui Zhang, Xie Chen, and Kai Yu. 2024. Vall-t: Decoder-only generative transducer for robust and decoding-controllable text-to-speech. *arXiv preprint arXiv:2401.14321*.

Zihao Fu, Wai Lam, Qian Yu, Anthony Man-Cho So, Shengding Hu, Zhiyuan Liu, and Nigel Collier. 2023. Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder. *arXiv preprint arXiv:2304.04052*.

Bing Han, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Yanming Qian, Yanqing Liu, Sheng Zhao, Jinyu Li, and Furu Wei. 2024. Vall-e r: Robust and efficient zero-shot text-to-speech synthesis via monotonic alignment. *arXiv preprint arXiv:2406.07855*.

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

Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *Proceedings of the ICASSP Conference 2020*, pages 7669–7673.

Shun Lei, Yixuan Zhou, Liyang Chen, Dan Luo, Zhiyong Wu, Xixin Wu, Shiyin Kang, Tao Jiang, Yahui Zhou, Yuxing Han, et al. 2024. Improving language model-based zero-shot text-to-speech synthesis with multi-scale acoustic prompts. In *Proceedings of the ICASSP Conference 2024*, pages 12662–12666.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of the DeeLIO in ACL Conference 2022*, pages 100–114.

Sheng Liu, Lei Xing, and James Zou. 2023. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Proceedings of the ICASSP Conference 2015*, pages 5206–5210.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. In *Proceedings of the Interspeech Conference 2020*, pages 2757–2761.

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

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2021. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. In *Proceedings of the Interspeech Conference 2021*, pages 2756–2760.

Yakun Song, Zhuo Chen, Xiaofei Wang, Ziyang Ma, and Xie Chen. 2024. Ella-v: Stable neural codec language modeling with alignment-guided sequence reordering. *arXiv preprint arXiv:2401.07333*.

Uenoyama Katsuya Aihara Tachibana, Hideyuki and Shunsuke. 2018. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *Proceedings of the ICASSP Conference 2018*, pages 4784–4788.

Cassia Valentini-Botinhao et al. 2017. Noisy speech database for training speech enhancement algorithms and tts models. *University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR)*.

Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2016. Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.

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

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. 2023b. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *Proceedings of the ICASSP Conference 2023*, pages 1–5.

Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. 2023c. Viola: Unified codec language models for speech recognition, synthesis, and translation. *arXiv preprint arXiv:2305.16107*.

Xiaofei Wang, Manthan Thakker, Zhuo Chen, Naoyuki Kanda, Sefik Emre Eskimez, Sanyuan Chen, Min Tang, Shujie Liu, Jinyu Li, and Takuya Yoshioka. 2024. Speechx: Neural codec language model as a versatile speech transformer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Detai Xin, Xu Tan, Kai Shen, Zeqian Ju, Dongchao Yang, Yuancheng Wang, Shinnosuke Takamichi, Hiroshi Saruwatari, Shujie Liu, Jinyu Li, et al. 2024. Rall-e: Robust codec language modeling with chain-of-thought prompting for text-to-speech synthesis. *arXiv preprint arXiv:2404.03204*.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. 2024. Zipformer: A faster and better encoder for automatic speech recognition. In *Proceedings of the ICLR Conference 2024*, pages 1–xx.

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

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *Proceedings of the Interspeech Conference 2019*, pages 1526–1530.

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