LLaSE-G1: Incentivizing Generalization Capability for LLaMA-based Speech Enhancement

Boyi Kang[1] *, Xinfa Zhu[1] *, Zihan Zhang[1] *, Zhen Ye[2], Mingshuai Liu[1], Ziqian Wang[1], Yike Zhu[1], Guobin Ma[1], Jun Chen[3], Longshuai Xiao[3], Chao Weng[3], Wei Xue[2], Lei Xie[1]
[1] Northwestern Polytechnical University
[2] The Hong Kong University of Science and Technology
[3] Huawei Technologies Co., Ltd.
{beaukang, xfzhu, zhzhang, lxie}@mail.nwpu.edu.cn

Abstract

Recent advancements in language models (LMs) have demonstrated strong capabilities in semantic understanding and contextual modeling, which have flourished in generative speech enhancement (SE). However, many LM-based SE approaches primarily focus on semantic information, often neglecting the critical role of acoustic information, which leads to acoustic inconsistency after enhancement and limited generalization across diverse SE tasks. In this paper, we introduce LLaSE-G1, a LLaMA-based language model that incentivizes generalization capabilities for speech enhancement. LLaSE-G1 offers the following key contributions: First, to mitigate acoustic inconsistency, LLaSE-G1 employs continuous representations from WavLM as input and predicts speech tokens from X-Codec2, maximizing acoustic preservation. Second, to promote generalization capability, LLaSE-G1 introduces dual-channel inputs and outputs, unifying multiple SE tasks without requiring task-specific IDs. Third, LLaSE-G1 outperforms prior task-specific discriminative and generative SE models, demonstrating scaling effects at test time and emerging capabilities for unseen SE tasks. Additionally, we release our code and models to support further research in this area [1].

1 Introduction

In recent years, large language models (LLMs) have made significant strides in natural language processing (NLP) (OpenAI, 2024), computer vision (CV) (Tschannen et al., 2024; Chang et al., 2022), and speech processing (Wang et al., 2023a; Zhang et al., 2023b), driving the rapid development of artificial intelligence technologies. In the NLP domain, LLMs have redefined text generation benchmarks through innovative pre-training and post-training paradigms, particularly excelling in few-shot and zero-shot learning scenarios. The impact of LLMs extends beyond unimodal textual processing. In CV research, integrating LLMs with visual models has sparked the rise of multimodal learning frameworks, facilitating more efficient processing of tasks such as image comprehension and generation. Similarly, the convergence of modalities is evident in the speech domain, where LLMs have enhanced the naturalness and accuracy of speech interaction systems. These advancements not only highlight the power of LLMs within individual domains but also underscore their potential for multimodal tasks.

Task Type	Distortion	Reference Signal
NS	Noise, Reverb	-
PLC	Noise, Packet Loss	Lossy Label
TSE	Noise, Reverb, Interfering Speech	Enrolled Speech
AEC	Noise, Reverb, Echo	Echo Speech
SS	Noise, Reverb, Interfering Speech	-

Table 1: Subtasks Definition in Speech Enhancement

As a fundamental task in the field of speech processing, *speech enhancement (SE)* aims to remove interference from noisy speech and separate and reconstruct clean target speech. Depending on the differences between the interfering speech and the target speech, sub-tasks can be defined as Noise Suppression (NS), Packet Loss Concealment (PLC), Target Speaker Extraction (TSE), Acoustic Echo Cancellation (AEC), Speech Separation (SS), and others, as detailed in Table 1. Neural SE models can generally be categorized into two types: discriminative (Zhao et al., 2024a,c) and generative (Wang et al., 2024). Deep learning-based discriminative SE models learn a mapping between degraded speech and the corresponding clean speech target. In contrast, generative SE models employ language models or diffusion models to learn the

*Equal contribution.
[1]LLaSE-G1 Codes and Demos

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13292–13305

data distribution of the target speech. Notable recent models, including SELM (Wang et al., 2024), TSELM (Tang et al., 2024), and GenSE (Yao et al., 2025), leverage semantic understanding and contextual modeling capabilities, achieving competitive performance in speech enhancement tasks. While traditional discriminative SE models require carefully designed architectures and task-specific loss functions, generative SE models offer a more flexible framework, enabling better scalability across different SE tasks.

Despite surpassing traditional discriminative models in speech quality, generative SE models still face challenges in acoustic preservation and task generalization. Many generative SE models rely on discrete speech tokens—typically extracted from speech codecs—as inputs to facilitate language modeling. However, as speech is inherently a continuous signal, using discrete tokens, especially semantic tokens, inevitably results in information loss (Yao et al., 2025), leading to acoustic inconsistencies after enhancement, such as changes in speaker timbre and intonation. Moreover, most generative models are focused on a single task, such as noise suppression, which limits their generalization across different SE tasks. Since SE tasks differ in their input, output, and underlying functions, it remains an open question whether LMs can serve as versatile, multi-task SE models.

In this paper, we argue that, with appropriate design, a single language model can be a powerful and versatile SE model. To this end, we propose LLaSE-G1, a LLaMA-based language model that incentivizes generalization capabilities across various SE tasks. The architecture of LLaSE-G1 is simple yet effective, consisting of a WavLM (Chen et al., 2022) encoder for feature extraction, a LLaMA-based language model for token prediction, and an X-codec2 (Ye et al., 2025) decoder for waveform reconstruction. Specifically, to address the acoustic inconsistency caused by the information loss inherent in discrete tokens, we replace the discrete token inputs with continuous representations extracted from the WavLM encoder and predict speech tokens obtained from X-codec2. The WavLM encoder provides sufficient speech details, and X-codec2 integrates semantic and acoustic features into speech tokens, thus maximizing acoustic preservation. Additionally, to incentivize generalization, LLaSE-G1 utilizes dual-channel inputs and outputs, unifying the degraded speech and optional reference signals and constraining all tasks under a cross-entropy loss function. Through extensive experiments, LLaSE-G1 demonstrates superior performance on NS, PLC, TSE, and AEC benchmarks. Furthermore, LLaSE-G1 exhibits emergent capabilities for previously unseen SE tasks, such as SS, and shows scaling effects at test time, where performance improves with increased compute.

In summary, our paper makes several key contributions:

- We propose LLaSE-G1, a LLaMA-based language model that incentivizes generalization capability for speech enhancement.
- We effectively address the acoustic inconsistency by leveraging both continuous and discrete representations, and we design dual-channel inputs and outputs, which unify various SE tasks without the need for task IDs. Notably, AEC, PLC, and SS tasks are being introduced to generative models for the first time.
- LLaSE-G1 outperforms existing models on several SE benchmarks and demonstrates scaling effects during test time and emergent capabilities for unseen SE tasks. We release the codes and checkpoints as open-source.

2 Related Work

Speech enhancement refers to the technology of recovering high-quality target speech from degraded speech, which includes multiple subtasks (Wang and Chen, 2018; Liu et al., 2022b). Traditional speech enhancement, which relies on statistical analysis and signal processing, often struggles with generalization in dynamic scenarios. With the development of deep learning, data-driven speech enhancement has become the mainstream approach and can be divided into two categories: discriminative SE and generative SE (Lemercier et al., 2023). Discriminative SE models learn a mapping between degraded speech and the corresponding clean speech targets, including methods such as time-frequency (T-F) masking (Williamson and Wang, 2017) and time-domain approaches (Luo and Mesgarani, 2018). In contrast, generative models reconstruct the clean speech by learning the data distribution of the target speech, such as diffusion-based generative models (Zhang et al., 2023a; Richter et al., 2023). Recently, researchers have also begun to explore the use of LMs to improve generative speech enhancement (Wang et al., 2024; Yao et al., 2025).

2.1 Discriminative Speech Enhancement

Traditionally, speech enhancement encompasses tasks such as NS, PLC, TSE, AEC, and SS, with NS also requiring dereverberation. For different tasks, discriminative models often have different architectures. In NS tasks, most models are based on the convolutional encoder-decoder (CED) architecture. FRCRN (Zhao et al., 2024a) adds a frequency recurrent network to the CED architecture, achieving excellent performance. In PLC tasks, Generative Adversarial Networks (GANs) are commonly used. BS-PLCNet (Zhang et al., 2024) uses multi-task learning and multi discriminators, winning the latest PLC Challenge (Diener et al., 2024). In TSE tasks, the speaker embedding paradigm is widely adopted. TSE approaches usually use speaker verification models (Desplanques et al., 2020; Wang et al., 2023b) to extract embeddings from enrollment audio and integrate into noise suppression networks. This approach has been successful in the personalized tracks of the Deep Noise Suppression challenges, as demonstrated by the TEA-PSE series models (Ju et al., 2023, 2022). For AEC tasks, an important issue is how to deal with the delay estimation and alignment between reference signals and microphone signals. DeepVQE (Indenbom et al., 2023a) utilizes attention-based delay estimation, employing fully neural networks to solve echo cancellation problems. For SS tasks, common models such as TF-GridNet (Wang et al., 2023c) and Mossformer2 (Zhao et al., 2024b) can only handle a fixed number of speakers. SepTDA (Lee et al., 2024) introduces a transformer decoder-based attractor, capable of handling a dynamic number of speakers, but still requires specifying the maximum number of speakers.

While these discriminative models have achieved excellent performance across various tasks, their generalization ability is limited by the availability of training data and model parameters (Welker et al., 2022). This can lead to performance degradation in unseen scenarios. Additionally, these models may introduce undesired speech distortion and phonetic inaccuracies to enhanced speech (Wang et al., 2020).

2.2 Generative Speech Enhancement

Early generative SE primarily relied on GANs and VAEs (Pascual et al., 2017; Fang et al., 2021). Although these approaches offered new perspectives, they still did not surpass the performance of discriminative models. In recent years, diffusion-based generative models have been applied to speech enhancement. CDiffusion (Lu et al., 2022) defines the conditional diffusion process by incorporating noisy data into the diffusion process. DiffSep (Scheibler et al., 2023) designs stochastic differential equations (SDE) (Song et al., 2021). By solving the corresponding reverse-time SDE, it is possible to recover individual sources from the mixture. Despite diffusion models achieving superior speech quality over discriminative models in noise suppression (NS) and source separation (SS), these tasks were previously independent of each other, requiring separate training of different models, and proving difficult to generalize to other tasks.

Recently, researchers have begun to explore the use of a joint framework, leveraging the capabilities of generative models to integrate multiple enhancement tasks into a single model. Nemo (Ku et al., 2024) and SpeechFlow (Liu et al., 2024) pretrained on large-scale datasets and can be adapted to downstream tasks such as NS and TSE through fine-tuning. AnyEnhance (Zhang et al., 2025) achieves both NS and TSE without the need for fine-tuning. It introduces a prompt guidance mechanism, enabling in-context learning capabilities.

With the rise of LMs in their ability to handle multiple tasks, LMs have also been introduced into speech enhancement. SELM (Wang et al., 2024) employs a WavLM-based k-means tokenizer and predicts clean tokens from noisy tokens, marking the first introduction of LMs into the NS domain. MaskSR (Li et al., 2024) uses a mask generation model to simultaneously handle noise, reverberation, clipping, and bandwidth limitation. However, existing unified enhancement models have not considered the echo cancellation task, which requires reference audio input and the need to address delay estimation and alignment issues. We suggest that by leveraging the powerful modeling capabilities of LMs, it is possible to develop a general speech enhancement model that unifies NS, PLC, TSE, AEC and SS.

3 LLaSE-G1

3.1 Overall Architecture

LLaSE-G1 is designed to incentivize generalization across various SE tasks with a single LLaMA-based LM (Grattafiori et al., 2024). As shown in

Figure 1, compared to previous specialist models such as FRCRN (Zhao et al., 2024a), TEA-PSE 3.0 (Ju et al., 2023), Align-ULCNet (Shetu et al., 2024b), mossformer2 (Zhao et al., 2024c) and BS-PLCNet (Zhang et al., 2024), LLaSE-G1 greatly simplifies the model structure, keeping three main components: (1) a WavLM encoder, (2) a LLaMA-based LM and (3) an X-codec2 decoder. Specifically, the WavLM encoder extracts continuous speech features from degraded speech. The LLaMA-based LM takes speech features as input and predicts discrete speech tokens extracted by X-codec2 in an autoregressive manner. Finally, the X-codec2 decoder reconstructs enhanced speech from predicted speech tokens.

Formally, let: 1. $\texttt{Vertorize}(X) = \{x_1, \ldots, x_N\}$ be the WavLM encoder, which converts input degraded speech X into N speech features. 2. $\texttt{Vertorize}(P) = \{p_1, \ldots, p_T\}$ be the WavLM encoder, which converts optional reference speech P into T speech features. 3. $\texttt{Tokenize}(Y) = \{y_1, \ldots, y_M\}$ be the X-codec2 encoder, which converts an enhanced speech Y into M speech tokens. 4. $\texttt{Detokenize}(\{y_1, \ldots, y_M\}) = \hat{Y}$ be the X-codec2 decoder, which reconstructs the waveform $\hat{Y}$ from its token representations. As the down-sampling rate of WavLM is the same as that of X-codec2, N is equal to M. Given a dataset $\mathcal{D} = \{(X_i, P_i, Y_i)\}_{i=1}^{S}$, where X_i is the degraded speech, Y_i is the enhanced speech and P_i is the reference speech or empty if unavailable, we represent each pair (X_i, P_i, Y_i) as a sequence $(x_1, \ldots, x_N, p_1, \ldots, p_T, y_1, \ldots, y_M)$. Since the X_i and P_i are always given as input during training and inference, we pad X_i and P_i to the same length and the LM θ focuses on learning the conditional probability:

$$\begin{aligned} &P(x_1, \ldots, x_N, p_1, \ldots, p_T, y_1, \ldots, y_M; \theta) \\ &= \prod_{j=1}^{M} P(y_j | x_1 \odot p_1, \ldots, x_j \odot p_j; \theta), \end{aligned} \tag{1}$$

where $\odot$ is the concatenation between x and p in the channel axis.

3.2 Maximizing Acoustic Preservation

As highlighted by WavChat (Ji et al., 2024), speech representations can be broadly categorized into two types: continuous and discrete representations. Continuous representations, typically extracted from self-supervised learning (SSL) models like HuBERT (Hsu et al., 2021) and WavLM (Chen et al., 2022), are considered lossless carriers for speech, capturing intricate speech details. In contrast, discrete representations are derived from speech codecs such as Encodec (Défossez et al., 2022) and DAC (Kumar et al., 2023), which, while facilitating language modeling, are lossy due to the information loss during quantization. To address this, LLaSE-G1 adopts continuous representations as input and predicts discrete representations, aiming to maximize acoustic preservation throughout the enhancement process.

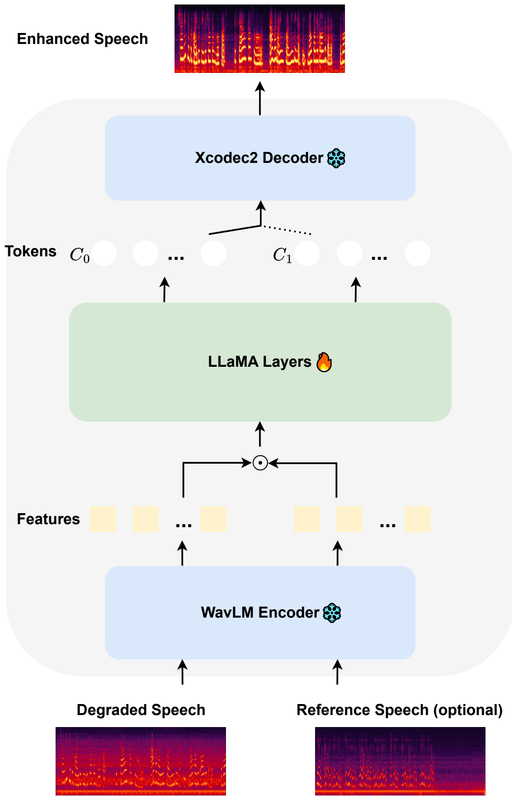


Figure 1: Overall Architecture of LLaSE-G1. LLaSE-G1 simplifies model architecture to support various SE tasks.

For continuous speech representations, we utilize WavLM as the extractor. WavLM is an SSL model that combines a convolutional feature encoder with a transformer encoder. Pre-trained on large-scale speech data, it excels across various speech-processing tasks. Previous research (Baas et al., 2023; Zhu et al., 2023) has shown that features extracted from the 6th layer of WavLM contain sufficient acoustic information for high-fidelity speech reconstruction. Therefore, we leverage the features from this layer as the input representations for the language model.

For discrete speech representations, we use X-Codec2 as the extractor. X-Codec2 is a recently developed speech codec that integrates semantic and acoustic features into a unified codebook, ensuring a 1D causal dependency. This design reflects the inherent left-to-right temporal structure of audio signals, while also preserving more acoustic information compared to traditional 1D semantic tokens. Consequently, we adopt the speech tokens extracted by X-Codec2 as the modeling target for the language model.

3.3 Unifying Various SE tasks

Although different speech enhancement tasks are applied in various scenarios, they share underlying commonalities, such as the need to determine which components should be removed from the noisy speech. To this end, LLaSE-G1 employs a dual-channel input and output framework that unifies several SE tasks within a single language model (LM). These tasks include NS, TSE, PLC, AEC, and SS.

Systematically, NS, PLC, and SS require only the degraded speech as input, while TSE and AEC need both the degraded speech and an additional reference speech. In contrast to previous PLC models, we do not use the lossy labels that indicate missing speech frames, thereby simplifying the data requirements.

To unify the input representations, we introduce a dual-channel input: one channel for the degraded speech and the other for the optional reference speech. These representations are padded to the same length and concatenated along the channel dimension. Notably, if the reference speech is unavailable, we set the second channel to zero.

While NS, PLC, AEC, and TSE tasks output a single enhanced speech, we note that AEC requires the removal of information related to the reference speech, while TSE necessitates the preservation of reference speech information. To address this, we introduce a dual-channel output with two linear projection heads to unify the output representations. The output embedding of the LLM is passed through these two linear projection to produce two embeddings from it. The first channel c_0 predicts tokens related to reference speech and the second channel c_1 predicts tokens irrelevant to reference speech. With these designs, for tasks including NS, AEC, and PLC, we employ a single-supervision strategy $\mathcal{L}_{\mathrm{S}}$ through the cross-entropy loss between c_0 and the tokens t_0 extracted from the clean signal:

$$\mathcal{L}_{\mathrm{S}} = -\frac{1}{N}\sum_{k=1}^{N} t_0^{(k)} \log\left(c_0^{(k)}\right) \tag{2}$$

For the TSE task, we implement a dual-supervision strategy $\mathcal{L}_{\mathrm{D}}$ with separate constraints for both outputs. The first output c_0 handles interfering speaker, while the second output c_1 is dedicated to target speaker extraction. The $\mathcal{L}_{\mathrm{D}}$ is formulated as:

$$\mathcal{L}_{\mathrm{D}} = -\frac{1}{N}\sum_{k=1}^{N} [t_0^{(k)} \log\left(c_0^{(k)}\right) + t_1^{(k)} \log\left(c_1^{(k)}\right)] \tag{3}$$

Importantly, in LLaSE-G1, SS is treated as an unseen task throughout the entire training process.

4 Experiments and Results

4.1 Experimental Setup

Datasets. For the training data, we use the Librispeech, HiFiTTS, and DNS Challenge datasets (Reddy et al., 2020; Dubey et al., 2023), along with internal datasets as original clean speech, totalling approximately 5000 hours. Room impulse responses (RIRs) are sourced from the DNS Challenge datasets. The noise data contains nearly 1000 hours, sourced from DEMAND, ESC-50, DNS Challenge, AEC Challenge (Cutler et al., 2023), and internal datasets.

Data augmentation. We utilized dynamic data augmentation during training. For the NS task, the clean audio and noise are randomly selected and mixed with a signal-to-noise Ratio (SNR) ranging from [-5,20] dB. Both clean and noisy signals have a 50% probability of adding reverberation. In the PLC task, we use a two-state first-order Markov chain to describe the packet loss status of the current frame and the next frame. The transition and hold probabilities for Markov states are selected between 0.05 and 0.95. We directly generate a binary mask sequence and apply it to the clean speech. For the AEC task, we randomly select a real echo signal and its corresponding reference signal from the far-end single talk in the AEC Challenge dataset. The signal-to-echo ratio (SER) ranges from -15 dB to 15 dB. Noise is added with a 20% chance, and the SNR is between -5 dB and 20 dB. For the Target Speaker Extraction (TSE) task, we select a clean speech segment and its corresponding auxiliary segment for the enrollment speech, while a different speaker is chosen for the interference speech. There is a 5% probability that no interfering speaker is present. Target speech and interfering speech are mixed with an SNR ranging from [-15, 15] dB, with an additional 10% probability of adding extra noise.

The audio length for each batch is 8 seconds. Before being fed into the model, the audio is randomly truncated to a length between 4 and 8 seconds to ensure the model's ability to generalize to different audio lengths. During training, the distribution of tasks (NS, PLC, AEC, and TSE) is evenly balanced. Within each batch, the data are of the same task type. Gradient accumulation is enabled to help the model adapt to multi-task learning, with parameter updates occurring every 20 steps.

Model configuration. We use the open-source checkpoints of WavLM-large[2] and X-codec2[3]. The LLaMA-based LM comprises 16 LLaMA layers, each with 16 attention heads, a dropout rate of 0.1, a hidden size of 2048, and an intermediate size of 4096. The total number of parameters in the model is approximately 1.07 billion. More details are given in Appendix A.1.

Baseline systems. We evaluate the performance of our LLaSE-G1 with several state-of-the-art (SOTA) models of each subtask, including the winners of the recent signal processing grand challenges (Reddy et al., 2020; Dubey et al., 2023; Cutler et al., 2023; Diener et al., 2024, 2022) for each task. Details of the baseline system and test set for each subtask are provided in Appendix A.

Evaluation Metrics. We use objective metrics to evaluate the performance of the baseline systems and our model. DNSMOS (Reddy et al., 2022) include speech quality (SIG), background noise quality (BAK), and overall quality (OVRL) of the audio. AECMOS (Purin et al., 2022) consists of echo annoyance MOS (EMOS) and other degradation MOS (DMOS). PLCMOS (Diener et al., 2023) is used to assess the quality of audio processed by PLC algorithms. All MOS scores range from 1 to 5, representing audio quality from low to high. SpeechBERTScore (SBS) (Saeki et al., 2024) is also employed to evaluate the semantic similarity between the enhanced audio and the reference audio. Following (Zhang et al., 2025), we use HuBERT-base[4] model to extract semantic features. For acoustic similarity, we calculate speaker similarity SimW_B based on the WavLM-base-sv model[5] to evaluate the performance. Subjective evaluations are also conducted using the Mean Opinion Score (MOS) as the primary metric to assess the model's performance.

Inference. For each task, we conduct single and multiple inferences. For multiple inferences, we infer 10 times and take the best result where the model's output is used as the input for the next inference. For the PLC task, we only employ the audio to be processed as input, without the lossy label. For the TSE task, we keep the enrollment audio unchanged during multiple inferences. For the AEC task, we only use the reference audio for the first inference, subsequent inferences are treated as NS tasks. For the SS task, we employ a two-stage inference strategy. First, we separate one speaker from the mixed audio, and then use the first separated speaker's audio as a reference for the second inference stage.

4.2 Experimental Results

4.2.1 Noise Suppression

Table 2 presents a comparison between the proposed LLaSE-G1 and several state-of-the-art (SOTA) discriminative and generative models. The "With Reverb" column corresponds to the test set containing reverberation, while the "No Reverb" column refers to the clean test set without reverberation. The results show that generative noise suppression (NS) models consistently outperform discriminative ones, particularly under reverberant conditions. Even with single inference, LLaSE-G1 surpasses most competing systems. When employing multiple inferences, its performance improves further, achieving a SOTA OVRL score of 3.49 on the "No Reverb" test set and 3.42 on the "With reverb" test set.

Notably, LLaSE-G1 demonstrates exceptional performance on the with reverb test set compared to other generative enhancement systems, further highlighting the efficiency of continuous representations and the Xcodec2 decoder in handling challenging noise suppression tasks.

4.2.2 Packet Loss Concealment

We compared LLaSE-G1 with the top-performing models (Zhang et al., 2024; Li et al., 2022; Liu et al., 2022a; Valin et al., 2022) from the most recent two challenges on the Interspeech 2022 PLC

[2] WavLM-Large on Hugging Face
[3] X-codec2 on Hugging Face
[4] Hubert-base on Hugging Face
[5] WavLM-base-sv on Huggingface

Table 2: DNSMOS scores on the Interspeech 2020 DNS Challenge blind test set. "D" represents Discriminative and "G" represents Generative. LLaSE-G1$_{\text{single}}$ and LLaSE-G1$_{\text{multi}}$ represent single inference and multiple inference using LLaSE-G1, respectively.

Model	Type	With Reverb			No Reverb		
		SIG	BAK	OVRL	SIG	BAK	OVRL
Noisy	-	1.76	1.50	1.39	3.39	2.62	2.48
Conv-TasNet	D	2.42	2.71	2.01	3.09	3.34	3.00
DEMUCS	D	2.86	3.90	2.55	3.58	4.15	3.35
FRCRN	D	2.93	2.92	2.28	3.58	4.13	3.34
SELM	G	3.16	3.58	2.70	3.51	4.10	3.26
MaskSR	G	3.53	4.07	3.25	3.59	4.12	3.34
AnyEnhance	G	3.50	4.04	3.20	3.64	4.18	3.42
GenSE	G	3.49	3.73	3.19	3.65	4.18	3.43
LLaSE-G1$_{\text{single}}$	G	3.59	4.10	3.33	3.66	4.17	3.42
LLaSE-G1$_{\text{multi}}$	G	**3.65**	**4.16**	**3.42**	**3.71**	**4.19**	**3.49**

blind test set (Diener et al., 2022). It is important to note that LLaSE-G1 operates as a blind PLC without the need for lossy labels. This means LLaSE-G1 autonomously determines whether to perform PLC without prior knowledge of which frames experienced packet loss, making it a more challenging task and distinct from the models participating in the PLC Challenge.

Table 3: DNSMOS OVRL and PLCMOS scores on ICASSP 2022 PLC-challenge blind testet.

Model	Type	OVRL	PLCMOS
Noisy	-	2.56	2.90
KuaishouNet	D	-	4.27
LPCNet	D	3.09	3.74
PLCNet	D	-	3.83
BS-PLCNet	D	3.20	4.29
LLaSE-G1$_{\text{single}}$	G	3.03	3.68
LLaSE-G1$_{\text{multi}}$	G	**3.27**	**4.30**

The results in Table 3 demonstrate significant improvement with our model through inference time scaling. Specifically, the multi-inference approach boosts both OVRL and PLCMOS scores, with OVRL increasing from 3.03 to 3.27 and PLCMOS rising from 3.68 to 4.30, highlighting its effectiveness. LLaSE-G1's results on blind PLC surpassed those of other models using informed PLC, demonstrating the powerful content understanding and generation capabilities of LMs.

4.2.3 Target Speaker Extraction

We use the ICASSP 2023 DNS blind test set (Dubey et al., 2023) for the TSE task evaluation, which includes two tracks: the headset track and the speakerphone track.

Table 4: pDNSMOS scores on ICASSP 2023 DNS-challenge blind testet.

Model	Type	Track 1			Track 2		
		SIG	BAK	OVRL	SIG	BAK	OVRL
Noisy	-	4.15	2.37	2.71	4.05	2.16	2.50
TEA-PSE 3.0	D	4.12	**4.05**	3.65	3.99	**3.95**	3.49
NAPSE	D	3.81	3.99	3.38	3.92	4.17	3.56
LLaSE-G1$_{\text{single}}$	G	**4.21**	3.99	**3.72**	4.08	3.84	3.55
LLaSE-G1$_{\text{multi}}$	G	4.20	3.97	3.70	**4.11**	3.86	**3.58**

As shown in Table 4, LLaSE-G1 consistently achieves significantly higher SIG MOS scores across both tracks, surpassing all other methods. This indicates that language model-based generative approaches offer higher audio quality with reduced signal distortion. While TEA-PSE 3.0 and NAPSE exhibit certain advantages on headset and speakerphone devices respectively, LLaSE-G1 delivers the best overall performance across both tracks, demonstrating superior device generalization compared to discriminative models.

4.2.4 Acoustic Echo Cancellation

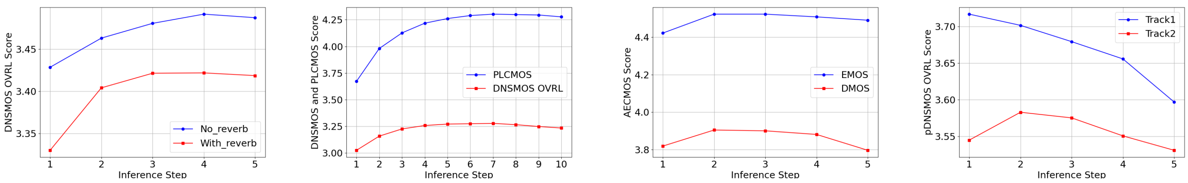
LLaSE-G1 is the first generative model to integrate the AEC task into a unified framework. As shown in Table 5, LLaSE-G1 demonstrates comparable performance to the SOTA discriminative AEC approaches, showcasing the potential of LMs-based generative models for the AEC task.

Table 5: AECMOS Echo (EMOS) and Degradation (DMOS) scores on ICASSP 2023 AEC-challenge blind test set."DT" represents double-talk, FEST means far-end only and NEST means near-end only.

Model	Type	DT		FEST	NEST
		EMOS	DMOS	EMOS	DMOS
Align-CRUSE	D	4.60	3.95	4.56	-
DeepVQE	D	**4.70**	**4.29**	4.69	**4.41**
ULCNetAENR	D	4.54	3.58	4.73	4.15
Align-ULCNet	D	4.60	3.80	**4.77**	4.28
LLaSE-G1$_{\text{single}}$	G	4.42	3.82	4.64	3.66
LLaSE-G1$_{\text{multi}}$	G	4.52	3.91	4.65	3.50

4.2.5 Emergent Capabilities and Scaling Effects at Test Time

Emergent capabilities. The SS task is not included in the training data, we use it to test the emergent capabilities of LLaSE-G1. When compared to other discriminative methods, our LLaMA-based LLaSE-G1 demonstrates significant emergent capabilities. With multiple inferences, our generative model outperforms discriminative meth-



(a) NS (b) PLC (c) AEC (d) TSE

Figure 2: Inference-time scaling results on different tasks

ods in OVRL scores of 3.17 and 3.25 on test sets, highlighting the potential of LLaSE-G1 to go beyond task-specific optimizations and adapt seamlessly to new tasks.

Table 6: DNSMOS scores on Libri2mix and WSJ0_2mix test set.

Model	Type	Libri2mix			WSJ0_2mix		
		SIG	BAK	OVRL	SIG	BAK	OVRL
Noisy	-	2.33	1.66	1.64	3.42	3.20	2.76
Sepformer	D	3.33	3.88	3.02	3.43	3.96	3.14
Mossformer2	D	3.44	3.94	3.11	3.50	4.05	3.23
LLaSE-G1$_{\text{single}}$	G	3.48	3.83	3.11	3.52	3.92	3.19
LLaSE-G1$_{\text{multi}}$	G	**3.50**	**3.90**	**3.17**	**3.55**	**3.97**	**3.25**

Inference-time scaling. As shown in Figure 2, scaling the inference time improves model performance across nearly all tasks. For the AEC and TSE tasks, performance peaks after the second inference, with EMOS improving from 4.42 to 4.52 and DMOS rising from 3.82 to 3.91. In contrast, the PLC task shows a significant performance boost with increased inference time, with PLCMOS rising from 3.67 to 4.30 and OVRL improving from 3.03 to 3.27, a gain of up to 25%. For the NS task, the OVRL score increases from 3.42 to 3.49 on the no_reverb dataset and from 3.33 to 3.42 on the with_reverb dataset. These results show that scaling test-time compute will initially improve performance, and decrease later due to the accumulation of acoustic distortion.

4.2.6 Semantic and Speaker Similarity

As shown in Table 7, we compare the semantic and speaker similarity between baseline systems and LLaSE-G1. Notably, TSE and AEC tasks are tested on the blind test sets where ground-truth speech is unavailable. So, we conduct evaluations of NS, PLC, and SS tasks. LLaSE-G1 outperforms generative SE models while getting slightly lower results in SBS, suggesting LLaSE-G1 effectively maintains speech content. Moreover, LLaSE-G1 achieves the highest SimW$_B$ in the NS task and competitive SimW$_B$ in the PLC and SS tasks, showing superior acoustic preservation capability.

Table 7: Semantic and speaker similarity results on various tasks, using the same test sets from previous subsections. As for the NS task, we report the average result across the No Reverb and With Reverb test sets.

Task	Model	Type	SBS	SimW$_B$
NS	FRCRN	D	**0.85**	0.980
	AnyEnhance	G	0.82	0.970
	SELM	G	0.72	0.965
	GenSE	G	0.78	0.974
	LLaSE-G1	G	0.83	**0.993**
PLC	BS-PLCNet	D	**0.95**	**0.999**
	LLaSE-G1	G	0.85	0.992
SS	Sepformer	D	0.85	0.980
	Mossformer2	D	**0.87**	**0.991**
	LLaSE-G1	G	0.82	0.988

4.2.7 Subjective Evaluation

We also conducted evaluations on subjective listening tests and user studies. The experimental setup and results are shown in Table 8. We primarily used the DNS blind test set for the subjective evaluation. A total of 12 participants were recruited, and each participant was asked to listen to 50 audio samples. The Mean Opinion Score (MOS) was calculated on a 5-point scale, ranging from 1 (poor quality) to 5 (excellent quality). Although GENSE achieves comparable performance to LLaSE-G1 on the DNSMOS metric, LLaSE-G1 still delivers better perceptual quality in subjective listening, further demonstrating its superior capability in preserving acoustic details.

Table 8: Subjective MOS comparison. “D” denotes discriminative models, “G” denotes generative models.

Model	Type	MOS
Clean	-	**4.8125**
FRCRN	D	3.8917
SELM	G	3.7750
GENSE	G	4.0083
LLaSE-G1	G	4.1208

Table 9: DNSMOS scores on DNS blind test set without reverb. "T" represents Discrete tokens, and "E" represents Embeddings. "S" represents Single inference, and "M" represents Multiple inference.

	Input	LM	Output	Inference	OVRL
Noisy	-	-	-	-	2.48
Baseline	WavLM Token	Transformer	HiFiGAN	S	3.26
	WavLM Embedding	Transformer	HiFiGAN	S	3.34
	WavLM Embedding	LLaMA	HiFiGAN	S	3.35
	Xcodec2 Token	LLaMA	X-codec2	S	3.29
Proposed	WavLM Embedding	LLaMA	X-codec2	S	3.43
	WavLM Embedding	LLaMA	X-codec2	M	**3.49**

4.2.8 Ablation Study

We conduct an ablation study to evaluate the effectiveness of input representations, output representations, and model backbone, choosing SELM as the baseline. As shown in Table 9, when replacing inputs and output with proposed continuous features and speech tokens, there is an obvious improvement, revealing the effectiveness of acoustic preservation. Additionally, Experimental results show that while Xcodec2 tokens are designed to better capture acoustic information, discrete tokens still discard more detailed content than continuous embeddings when used as model inputs. Besides, there is no performance drop when replacing the full attention Transformer with casual attention LLaMA layers. Finally, adopting a multi-inference strategy further boosts performance.

5 Conclusion

In this study, we propose LLaSE-G1, a general LLaMA-based framework that unifies a wide range of speech enhancement (SE) tasks. Specifically, we utilize continuous acoustic features as input and predict 1D speech tokens to maximize fidelity to the original audio. To support multiple SE tasks, we design dual-channel inputs and outputs within a unified architecture. Extensive experiments demonstrate that LLaSE-G1 achieves state-of-the-art performance across various benchmarks, establishing it as a strong foundation model. Furthermore, LLaSE-G1 exhibits clear test-time scaling behavior and emerging generalization capabilities to previously unseen SE tasks. Notably, our framework requires no additional prompts to differentiate between SE tasks, instead relying on the LLM's intrinsic ability to infer the task type from the input itself.

Limitations

Although LLaSE-G1 demonstrates promising results across diverse SE tasks, there are several limitations that can be addressed towards LLaSE-G2. First, LLaSE-G1 operates at a 16,000 Hz sampling rate due to WavLM and X-codec2. We plan to support full-band audio and super-resolution generation in future research. Second, the training data and model size of LLaSE-G1 are relatively small as compared with that used in mainstream audio langauge models for understanding and conversation tasks. Hence we would like to further scale up data and model size to boost performance in generative speech enhancement.

References

Matthew Baas, Benjamin van Niekerk, and Herman Kamper. 2023. Voice conversion with just nearest neighbors. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 2053–2057. ISCA.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. 2022. Maskgit: Masked generative image transformer. *Preprint*, arXiv:2202.04200.

Sanyuan Chen, Wang, and ... 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.

Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. 2020. Librimix: An open-source dataset for generalizable speech separation. *Preprint*, arXiv:2005.11262.

Ross Cutler, Ando Saabas, and ... 2023. Icassp 2023 acoustic echo cancellation challenge. *Preprint*, arXiv:2309.12553.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 3830–3834. ISCA.

Lorenz Diener, Solomiya Branets, Ando Saabas, and Ross Cutler. 2024. The icassp 2024 audio deep packet loss concealment challenge. *Preprint*, arXiv:2402.16927.

Lorenz Diener, Marju Purin, Sten Sootla, Ando Saabas, Robert Aichner, and Ross Cutler. 2023. Plcmos – a data-driven non-intrusive metric for the evaluation of packet loss concealment algorithms. *Preprint*, arXiv:2305.15127.

Lorenz Diener, Sten Sootla, Solomiya Branets, Ando Saabas, Robert Aichner, and Ross Cutler. 2022. Interspeech 2022 audio deep packet loss concealment challenge. *Preprint*, arXiv:2204.05222.

Harishchandra Dubey, Ashkan Aazami, and ... 2023. Icassp 2023 deep noise suppression challenge. *Preprint*, arXiv:2303.11510.

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

Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. 2019. Demucs: Deep extractor for music sources with extra unlabeled data remixed. *Preprint*, arXiv:1909.01174.

Huajian Fang, Guillaume Carbajal, Stefan Wermter, and Timo Gerkmann. 2021. Variational autoencoder for speech enhancement with a noise-aware encoder. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Aaron Grattafiori, Abhimanyu Dubey, and ... 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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. *Preprint*, arXiv:2106.07447.

Evgenii Indenbom, Nicolae-Catalin Ristea, Ando Saabas, Tanel Parnamaa, Jegor Guzvin, and Ross Cutler. 2023a. Deepvqe: Real time deep voice quality enhancement for joint acoustic echo cancellation, noise suppression and dereverberation. *Preprint*, arXiv:2306.03177.

Evgenii Indenbom, Nicolae-Cătălin Ristea, Ando Saabas, Tanel Pärnamaa, and Jegor Gužvin. 2023b. Deep model with built-in cross-attention alignment for acoustic echo cancellation. *Preprint*, arXiv:2208.11308.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Yidi Jiang, Jingzhen He, Yunfei Chu, Jin Xu, and Zhou Zhao. 2024. Wavchat: A survey of spoken dialogue models. *Preprint*, arXiv:2411.13577.

Yukai Ju, Jun Chen, Shimin Zhang, Shulin He, Wei Rao, Weixin Zhu, Yannan Wang, Tao Yu, and Shidong Shang. 2023. Tea-pse 3.0: Tencent-ethereal-audio-lab personalized speech enhancement system for icassp 2023 dns challenge. *Preprint*, arXiv:2303.07704.

Yukai Ju, Wei Rao, Xiaopeng Yan, Yihui Fu, Shubo Lv, Luyao Cheng, Yannan Wang, Lei Xie, and Shidong Shang. 2022. TEA-PSE: tencent-ethereal-audio-lab personalized speech enhancement system for

ICASSP 2022 DNS challenge. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 9291–9295. IEEE.

Pin-Jui Ku, Alexander H. Liu, Roman Korostik, Sung-Feng Huang, Szu-Wei Fu, and Ante Jukic. 2024. Generative speech foundation model pretraining for high-quality speech extraction and restoration. *CoRR*, abs/2409.16117.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved rvqgan. *Preprint*, arXiv:2306.06546.

Younglo Lee, Shukjae Choi, Byeong-Yeol Kim, Zhongqiu Wang, and Shinji Watanabe. 2024. Boosting unknown-number speaker separation with transformer decoder-based attractor. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 446–450. IEEE.

Jean-Marie Lemercier, Julius Richter, Simon Welker, and Timo Gerkmann. 2023. Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Nan Li, Xiguang Zheng, Chen Zhang, Liang Guo, and Bing Yu. 2022. End-to-end multi-loss training for low delay packet loss concealment. In *Interspeech 2022*, pages 585–589.

Xu Li, Qirui Wang, and Xiaoyu Liu. 2024. Masksr: Masked language model for full-band speech restoration. *Preprint*, arXiv:2406.02092.

Alexander H. Liu, Matthew Le, Apoorv Vyas, Bowen Shi, Andros Tjandra, and Wei-Ning Hsu. 2024. Generative pre-training for speech with flow matching. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

B. Liu, Q. Song, M. Yang, W. Yuan, and T. Wang. 2022a. Plcnet: Realtime packet loss concealment with semi-supervised generative adversarial network. In *Interspeech*, pages 575–579.

Haohe Liu, Xubo Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. 2022b. Voicefixer: A unified framework for high-fidelity speech restoration. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 4232–4236. ISCA.

Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao. 2022. Conditional diffusion probabilistic model for speech enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7402–7406. IEEE.

Yi Luo and Nima Mesgarani. 2018. Real-time single-channel dereverberation and separation with time-domain audio separation network. In *19th Annual Conference of the International Speech Communication Association, Interspeech 2018, Hyderabad, India, September 2-6, 2018*, pages 342–346. ISCA.

Yi Luo and Nima Mesgarani. 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Santiago Pascual, Antonio Bonafonte, and Joan Serrà. 2017. Segan: Speech enhancement generative adversarial network. *Preprint*, arXiv:1703.09452.

Marju Purin, Sten Sootla, Mateja Sponza, Ando Saabas, and Ross Cutler. 2022. Aecmos: A speech quality assessment metric for echo impairment. *Preprint*, arXiv:2110.03010.

Chandan K. A. Reddy, Vishak Gopal, and ... 2020. The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. *Preprint*, arXiv:2005.13981.

Chandan K A Reddy, Vishak Gopal, and Ross Cutler. 2022. Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. *Preprint*, arXiv:2110.01763.

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. 2023. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2351–2364.

Takaaki Saeki, Soumi Maiti, Shinnosuke Takamichi, Shinji Watanabe, and Hiroshi Saruwatari. 2024. Speechbertscore: Reference-aware automatic evaluation of speech generation leveraging nlp evaluation metrics. *Preprint*, arXiv:2401.16812.

Robin Scheibler, Youna Ji, Soo-Whan Chung, Jaeuk Byun, Soyeon Choe, and Min-Seok Choi. 2023. Diffusion-based generative speech source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Shrishti Saha Shetu, Naveen Kumar Desiraju, Jose Miguel Martinez Aponte, Emanuël A. P. Habets, and Edwin Mabande. 2024a. A hybrid approach for low-complexity joint acoustic echo and noise reduction. *Preprint*, arXiv:2408.15746.

Shrishti Saha Shetu, Naveen Kumar Desiraju, Wolfgang Mack, and Emanuël A. P. Habets. 2024b. Align-ulcnet: Towards low-complexity and robust acoustic echo and noise reduction. *Preprint*, arXiv:2410.13620.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. 2021. Attention is all you need in speech separation. *Preprint*, arXiv:2010.13154.

Beilong Tang, Bang Zeng, and Ming Li. 2024. Tselm: Target speaker extraction using discrete tokens and language models. *Preprint*, arXiv:2409.07841.

Michael Tschannen, Cian Eastwood, and Fabian Mentzer. 2024. Givt: Generative infinite-vocabulary transformers. *Preprint*, arXiv:2312.02116.

Jean-Marc Valin, Ahmed Mustafa, Christopher Montgomery, Timothy B. Terriberry, Michael Klingbeil, Paris Smaragdis, and Arvindh Krishnaswamy. 2022. Real-time packet loss concealment with mixed generative and predictive model. *Preprint*, arXiv:2205.05785.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.

DeLiang Wang and Jitong Chen. 2018. Supervised speech separation based on deep learning: An overview. *IEEE ACM Trans. Audio Speech Lang. Process.*, 26(10):1702–1726.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Peidong Wang, Ke Tan, and DeLiang Wang. 2020. Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:39–48.

Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe. 2023c. Tf-gridnet: Integrating full- and sub-band modeling for speech separation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:3221–3236.

Ziqian Wang, Xinfa Zhu, Zihan Zhang, YuanJun Lv, Ning Jiang, Guoqing Zhao, and Lei Xie. 2024. Selm: Speech enhancement using discrete tokens and language models. *Preprint*, arXiv:2312.09747.

Simon Welker, Julius Richter, and Timo Gerkmann. 2022. Speech enhancement with score-based generative models in the complex STFT domain. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2928–2932. ISCA.

Donald S. Williamson and DeLiang Wang. 2017. Time-frequency masking in the complex domain for speech dereverberation and denoising. *IEEE ACM Trans. Audio Speech Lang. Process.*, 25(7):1492–1501.

Xiaopeng Yan, Yindi Yang, Zhihao Guo, Liangliang Peng, and Lei Xie. 2023. The npu-elevoc personalized speech enhancement system for icassp2023 dns challenge. *Preprint*, arXiv:2303.06811.

Jixun Yao, Hexin Liu, Chen Chen, Yuchen Hu, Eng-Siong Chng, and Lei Xie. 2025. Gense: Generative speech enhancement via language models using hierarchical modeling. *Preprint*, arXiv:2502.02942.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi DAI, Hongzhan Lin, Jianyi Chen, Xingjian Du, Liumeng Xue, Yunlin Chen, Zhifei Li, Lei Xie, Qiuqiang Kong, Yike Guo, and Wei Xue. 2025. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *Preprint*, arXiv:2502.04128.

Chenshuang Zhang, Chaoning Zhang, Sheng Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In So Kweon. 2023a. A survey on audio diffusion models: Text to speech synthesis and enhancement in generative AI. *CoRR*, abs/2303.13336.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023b. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *Preprint*, arXiv:2305.11000.

Junan Zhang, Jing Yang, Zihao Fang, Yuancheng Wang, Zehua Zhang, Zhuo Wang, Fan Fan, and Zhizheng Wu. 2025. Anyenhance: A unified generative model with prompt-guidance and self-critic for voice enhancement. *Preprint*, arXiv:2501.15417.

Zihan Zhang, Jiayao Sun, Xianjun Xia, Chuanzeng Huang, Yijian Xiao, and Lei Xie. 2024. Bsplcnet: Band-split packet loss concealment network with multi-task learning framework and multi-discriminators. *Preprint*, arXiv:2401.03687.

Shengkui Zhao, Bin Ma, Karn N. Watcharasupat, and Woon-Seng Gan. 2024a. Frcrn: Boosting feature representation using frequency recurrence for monaural speech enhancement. *Preprint*, arXiv:2206.07293.

Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jia Qi Yip, Dianwen Ng, and Bin Ma. 2024b. Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 10356–10360. IEEE.

Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jiaqi Yip, Dianwen Ng, and Bin Ma. 2024c. Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation. *Preprint*, arXiv:2312.11825.

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. *CoRR*, abs/2310.07246.

A Appendix for Experimental Set Up

A.1 Model Configuration

We use the open-source checkpoints of WavLM-large and X-codec2. The LLaMA-based LM comprises 16 LLaMA layers, each with 16 attention heads, a dropout rate of 0.1, a hidden size of 2048, and an intermediate size of 4096. The total number of parameters in the model is approximately 1.07 billion, which includes all learnable weights and biases across all layers. The model has 2 input linear layers and 2 output linear layers. The input layer maps the 1024-dimensional WavLM embedding to another 1024-dimensional vector, while the output layer transforms a 2048-dimensional vector into a 65536-dimensional vector, which is the codebook size of Xcodec2.

We trained the model for 100,000 steps using 4 NVIDIA L40 GPUs, with a batch size of 6 per GPU and the AdamW optimizer. The learning rate is set to 1e-4.

A.2 Test Sets

NS: Interspeech 2020 DNS Challenge blind Test Set. (Reddy et al., 2020) It contains 600 clips (300 synthetic and 300 real), with synthetic clips generated using clean speech and noise not seen during training, and real clips crowdsourced in diverse noisy conditions.

PLC: Interspeech 2022 PLC Challenge test set (Diener et al., 2022) This is a realistic evaluation dataset based on packet loss patterns from actual calls, providing a methodology for comparing different approaches and a new objective metric to help researchers improve their techniques.

TSE: ICASSP 2023 DNS Challenge blind Test Set (Dubey et al., 2023) The blind test set includes two tracks Headset and Speakerphone with clips featuring 10-30 seconds of enrollment speech, with or without noise. It is used for final rankings and evaluates both personalized and non-personalized models using the Personalized ITU-T P.835 framework.

AEC: ICASSP 2023 AEC Challenge blind Test Set (Cutler et al., 2023) The blind test set in the AEC Challenge consists of real-world data collected from over 10,000 diverse audio devices and environments. It is used to determine the final competition winners. The dataset includes recordings of both single-talk and double-talk scenarios, with varying conditions like background noise, reverberation, and device distortions.

SS: Libri2mix (Cosentino et al., 2020), WSJ0-2mix. These two test sets are commonly used in speech separation, which is mixed from librispeech and WSJ datasets.

A.3 Baseline Systems

NS: For discriminative systems, we choose Conv-TasNet (Luo and Mesgarani, 2019),DEMUCS (Défossez et al., 2019),FRCRN (Zhao et al., 2024a) , which is recent SOTA models on noise suppression. For generative systems, we choose SELM (Wang et al., 2024), which introduce LM to speech enhancement, and GenSE (Yao et al., 2025) and AnyEnhance (Zhang et al., 2025), 2 newly released SOTA-level generative speech enhancement systems.

PLC: we use BS-PLCNet (Zhang et al., 2024), Team Kuaishow (Li et al., 2022), which are the winners of the 2024 challenge and 2022 challenge respectively, and other systems in challenge like PLCNet (Liu et al., 2022a) and LPCNet (Valin et al., 2022) as our baseline systems.

TSE: We compare our model with two baseline systems: TEA-PSE 3.0 (Ju et al., 2023), the winner of the challenge, and NAPSE (Yan et al., 2023), which placed second.

AEC: For baseline comparison, we choose recent efficient and state-of-the-art systems as our baseline systems, including UCLNet, (Shetu et al., 2024a),AlignUCLNet (Shetu et al., 2024b),AlignCruse (Indenbom et al., 2023b) and DeepVQE (Indenbom et al., 2023a), which is the state-of-the-art model in the AEC task.

SS: We use SOTA discriminative speech separation systems such as Sepformer (Subakan et al., 2021) and Mossformer2 (Zhao et al., 2024c), which is the SOTA system on speech separation, as our SS baseline systems.