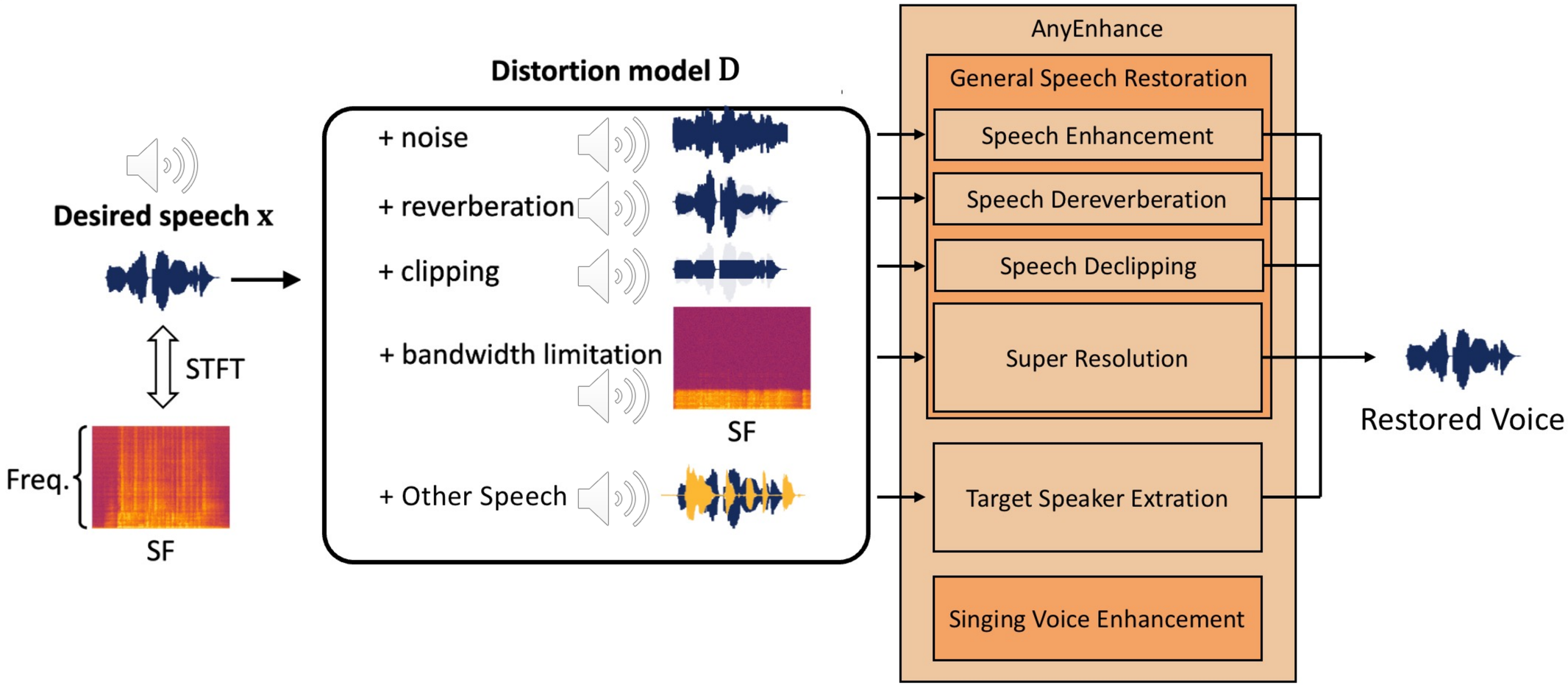
香港中文大學(深圳) | 数据科学学院
The Chinese University of Hong Kong, Shenzhen | School of Data Science

AnyEnhance: A Unified Generative Model with Prompt-Guidance and Self-Critic for Voice Enhancement

Junan Zhang
CUHKSZ

- **https://github.com/viewfinder-annn**
- **first-year phd student, advised by prof. zhizheng wu, bachelor in fudan university**
  - **interest: audio/music generation, generative models**
- **collaboration welcome!**

Background



Speech Enhancement Model in the Era of Large Generative Models:

- Multi-Task, Multi-Domain Pretraining
- In-Context-Learning

Background

Category	Model	Enhancement Tasks					Simultaneity	Domain	Sampling Rate	Supports Speaker Prompt?	Fine-tuning Required?
		Denoise	Dereverb	Declip	SR	TSE					
Multi-Task	Voicefixer[1]	✓	✓	✓	✓		✓	Speech	44.1kHz	✗	No
	MaskSR[2]	✓	✓	✓	✓		✓	Speech	44.1kHz	✗	No
	SpeechX[3]	✓				✓		Speech	16kHz	✗	No
	SpeechFlow[4]	✓				✓	✓	Speech	16kHz	✗	Yes
	NeMo[5]	✓				✓		Speech	16kHz	✗	Yes
	Uniaudio[6]	✓	✓			✓		Speech	16kHz	✗	Yes
Multi-Domain	AudioSR[7]				✓			Speech, Singing	48kHz	✗	No
Multi-Task Multi-Domain	AnyEnhance	✓	✓	✓	✓	✓	✓	Speech, Singing	44.1kHz	✓	No

Limitation in existing methods:

1. Task and Domain Universallity/Simutaneity
2. Prompt audio support

AnyEnhance:

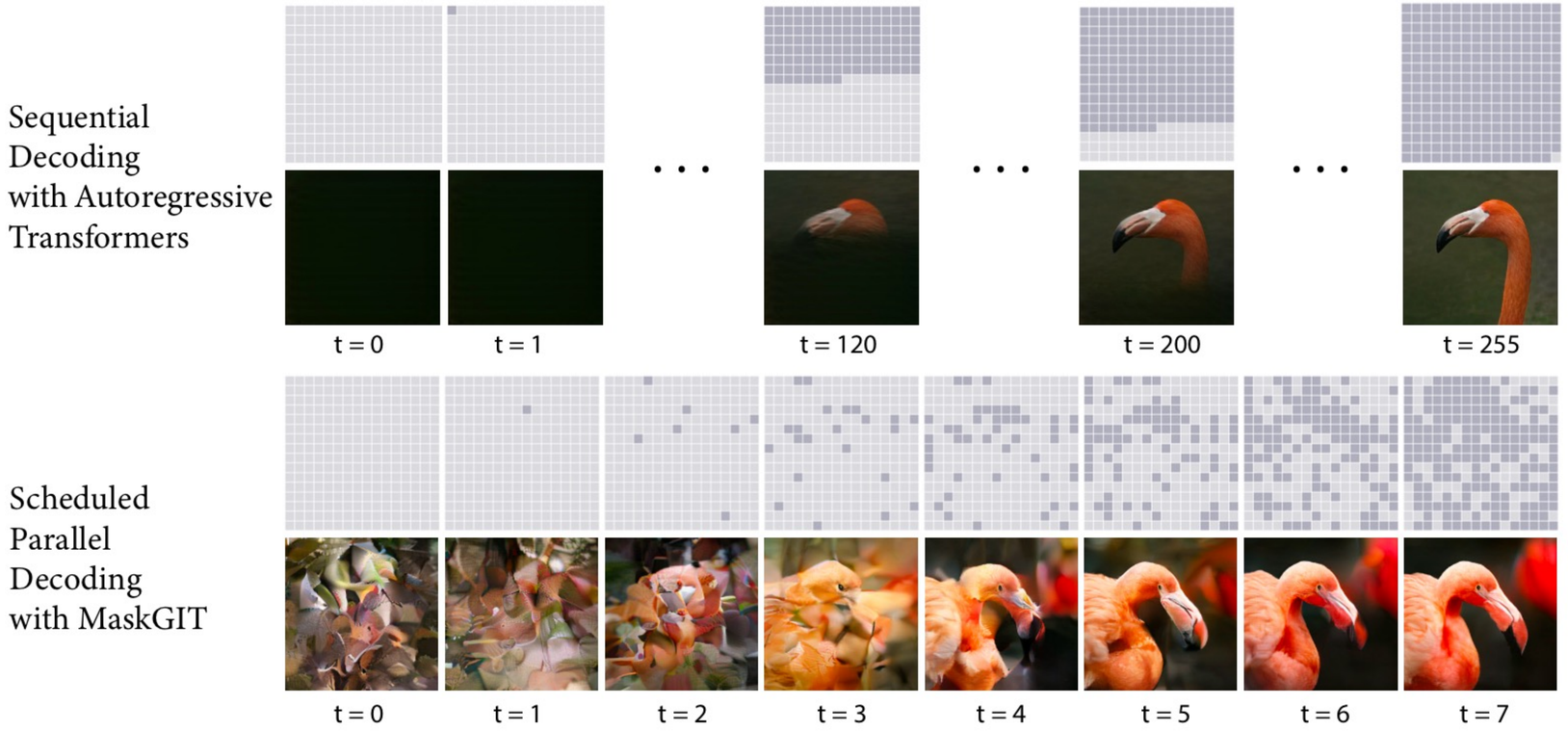
1. Multi-Task, Multi-Domain, Simultaneous
2. Supports prompt audio

What's more:

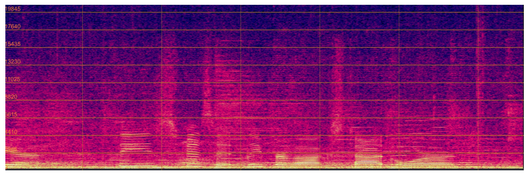
1. Self-critic for more stable sampling

Background

masked generative models

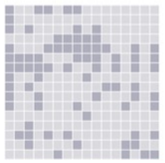


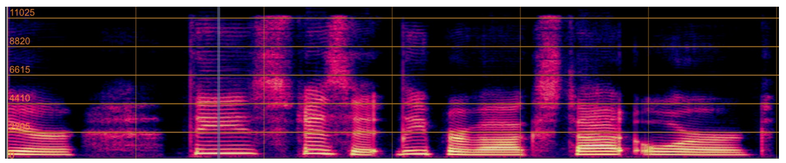
input



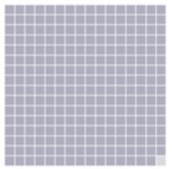
t=5

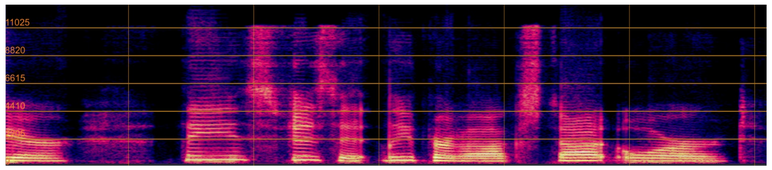
t=50





t=100





Background

Distortion Type	Mathematical Definition	Description
Noise	$d_{\text{noise}}(y) = y + n$	n is the noise signal.
Reverb	$d_{\text{reverb}}(y) = y * r$	r is the room impulse response.
Clipping	$d_{\text{clip}}(y) = \max(\min(y, \gamma), -\gamma)$	$\gamma \in [0, 1]$ is the clipping threshold.
Bandwidth Limitation	$d_{\text{bw}}(y) = \text{Resample}(y, \text{freq})$	freq is the target lower sampling frequency.
Other Voice	$d_{\text{voice}}(y) = y + v$	v is the other voice signal.

Previous works[1], [18] model distortions in a sequential order. Similarly, we model the overall distortion $D(\cdot)$ as a **composite function**:

$$D(x) = d_1 \circ d_2 \circ \dots d_Q(x),\ d_q \in \mathbb{D},\ q = 1, 2, \dots, Q,$$

Previous work: always adds noise last

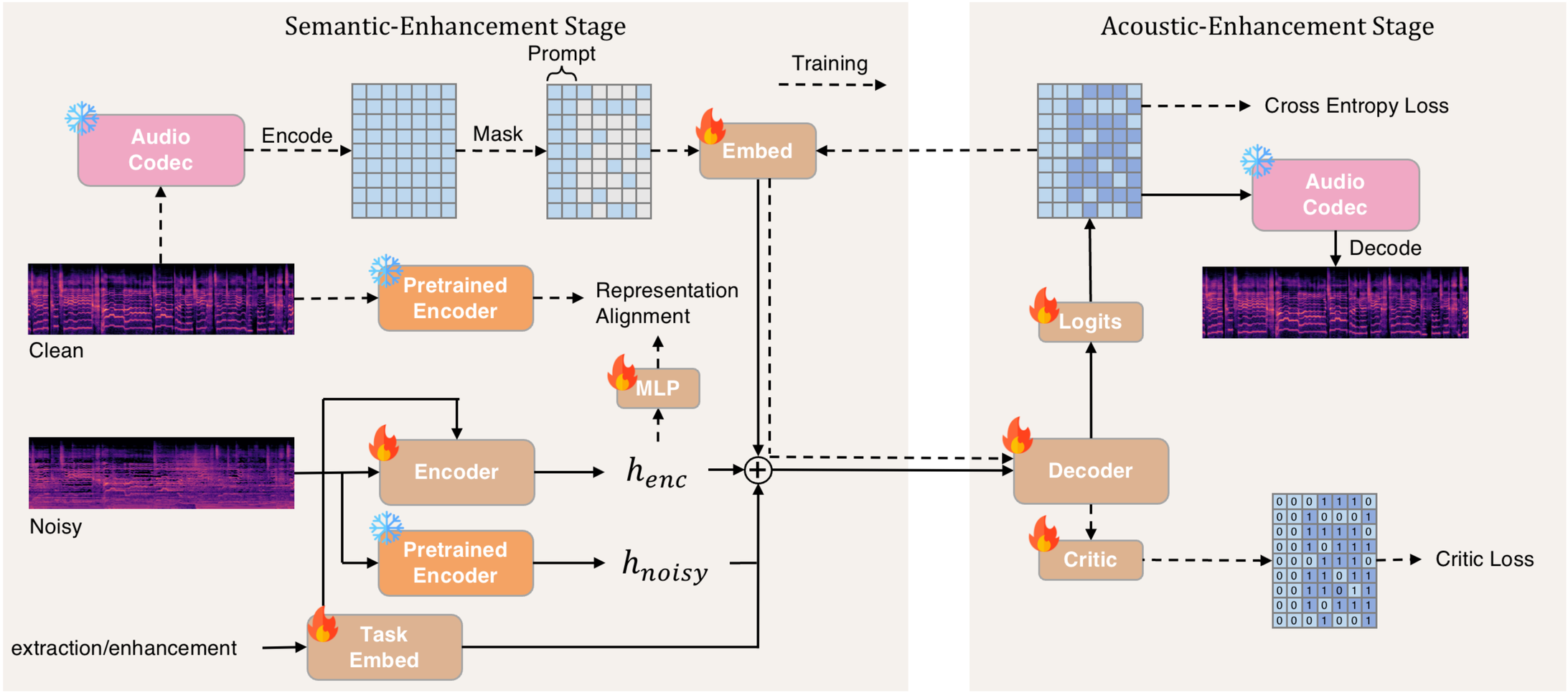
$$d_{\text{noise}} \circ d_{\text{bw}} \circ d_{\text{clip}} \circ d_{\text{reverb}}$$

Ours: noise is added before reverberation, and more realistic vocal effect

$$D(x) = d_{\text{bw}} \circ d_{\text{clip}} \circ d_{\text{reverb}} \circ d_{\text{noise}} \circ d_{\text{vocal\_effect}}$$

Where $d_{\text{vocal\_effect}} = d_{\text{reverb}} \circ d_{\text{eq}}$ is the vocal effect chain.

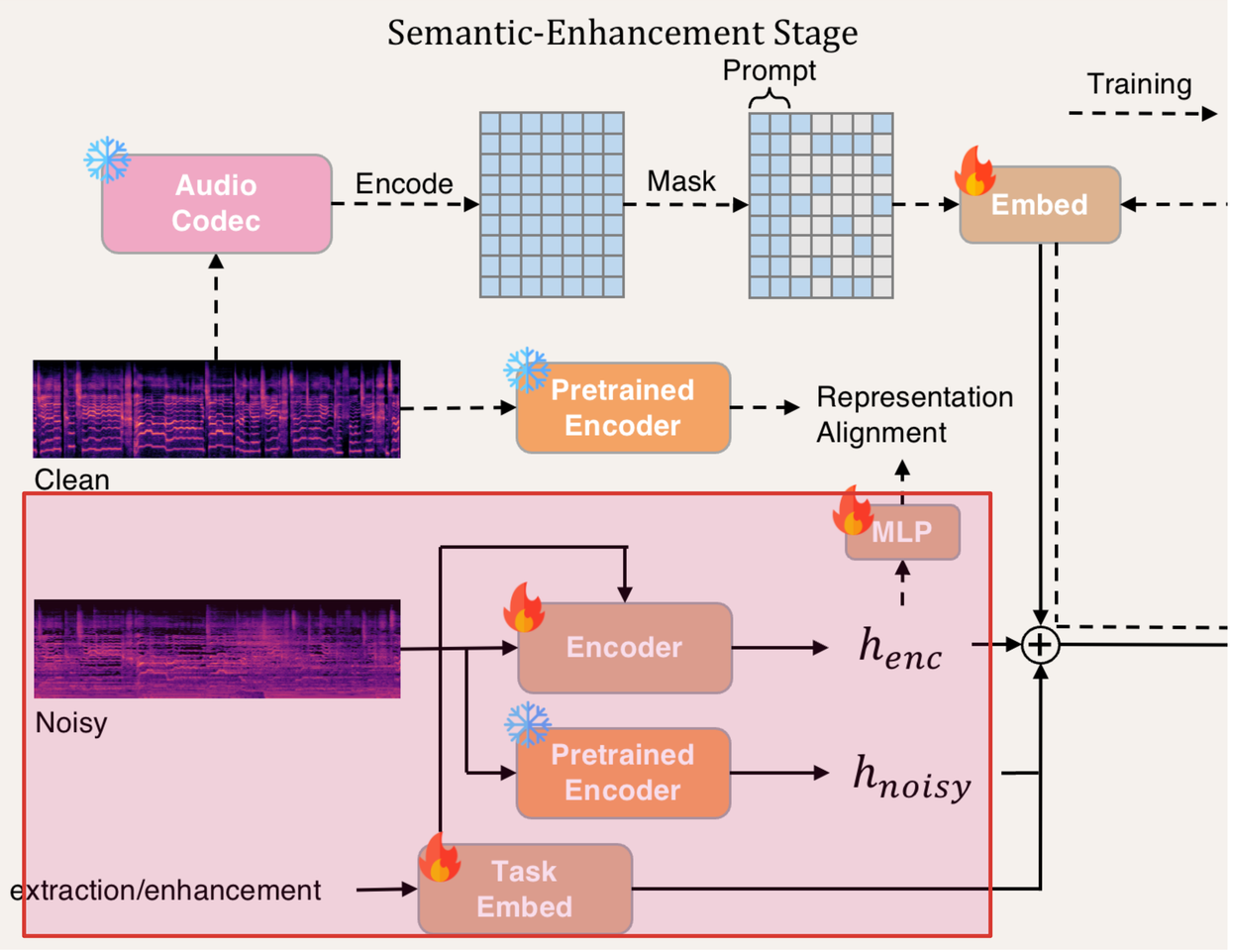
AnyEnhance



2 stages:

1. semantic enhancement: noisy audio -> semantic features
2. acoustic enhancememt: semantic features + masked audio codes -> full audio codes

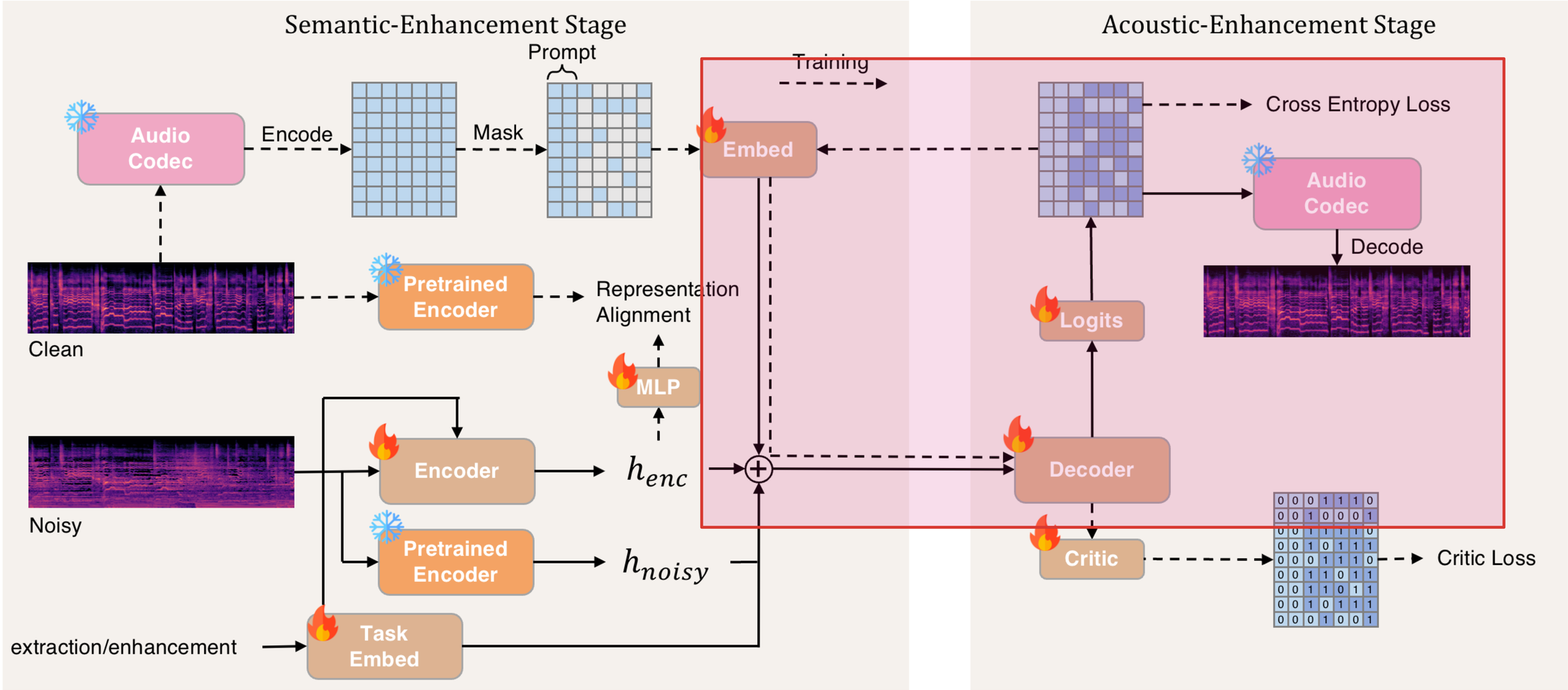
AnyEnhance



1. <u>semantic enhancement: noisy audio -> semantic features</u>
2. acoustic enhancememt: semantic features + masked audio codes -> full audio codes

$$\mathcal{L}_{\text{REPA}} = \mathbb{E}_{\mathbf{x}} \left[-\sum_{t=1}^{T} \text{sim}\left(\mathbf{y}^{[t]}, h_{\phi}(h_{\text{enc}}^{[t]}) \right) \right],$$

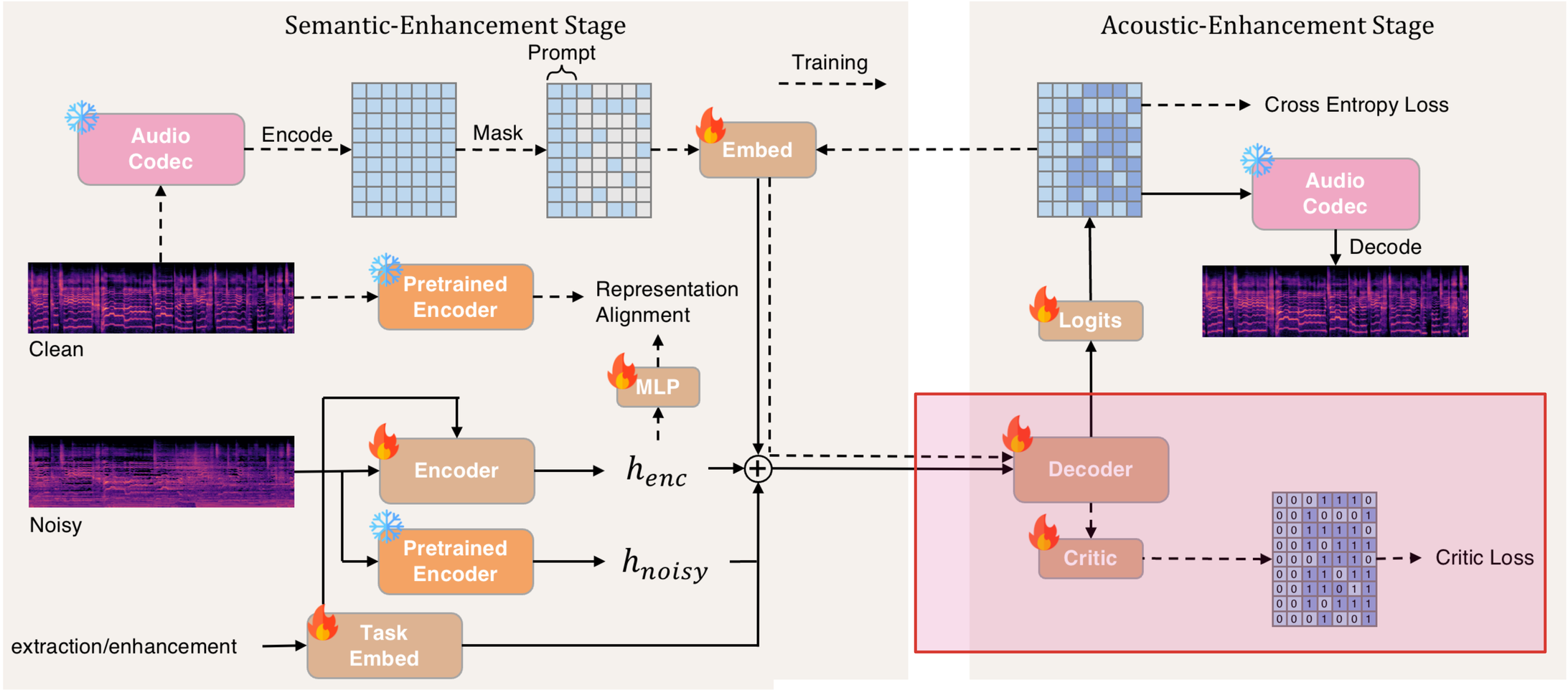
AnyEnhance



1. semantic enhancement: noisy audio -> semantic features
2. acoustic enhancememt: semantic features + masked audio codes -> full audio codes

$$\mathcal{L}_{\text{mask}} = \mathbb{E}_{\mathbf{X}\in\mathcal{D},t\in[0,T]}\left[-\sum_{i=1}^{N} m_{t,i}\cdot\log\left(p_\theta(x_i \mid \mathbf{X}_t, \mathbf{C})\right)\right].$$

AnyEnhance



Self-Critic:

used to tell whether a token is real

during sample: we use critic head's output instead of original token's logits as confidence score

$$\mathcal{L}_{\text{critic}} = \mathbb{E}_{\substack{\mathbf{X}\in\mathcal{D},t\in[0,T],\\ p_\theta(\tilde{\mathbf{X}}_0|\mathbf{X}_t,\mathbf{C})}} \left[-\sum_{i=1}^{N} m_{t,i} \cdot \log\left(p_\phi(\hat{m}_{t,i} \mid \tilde{\mathbf{X}}_0, \mathbf{C})\right) \right].$$

$$\mathcal{L} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{REPA}} + \mathcal{L}_{\text{critic}}.$$

AnyEnhance

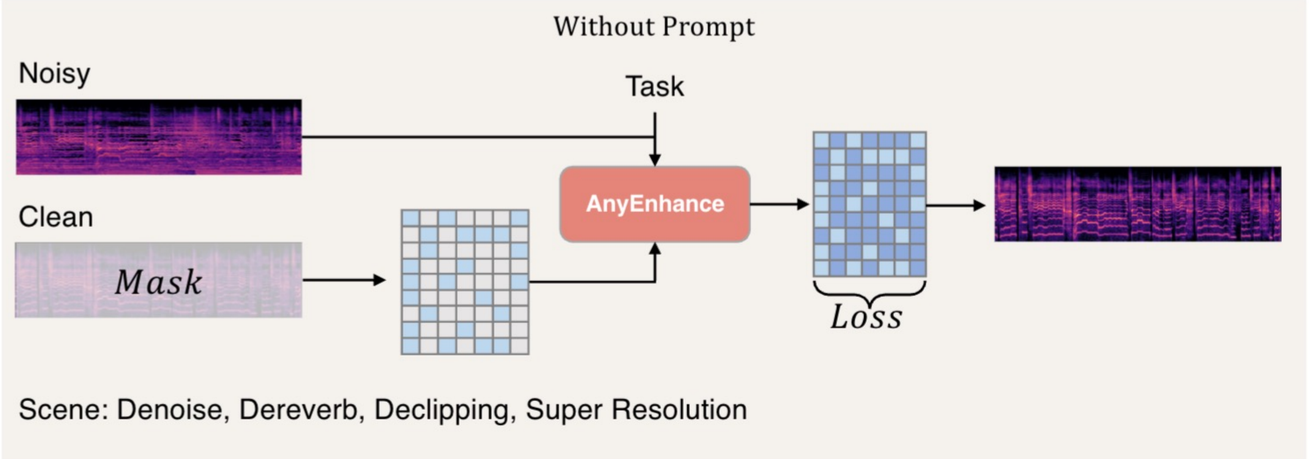
Inspired by zero-shot tts, during training, we preserve a portion of clean audio with a certain probability.

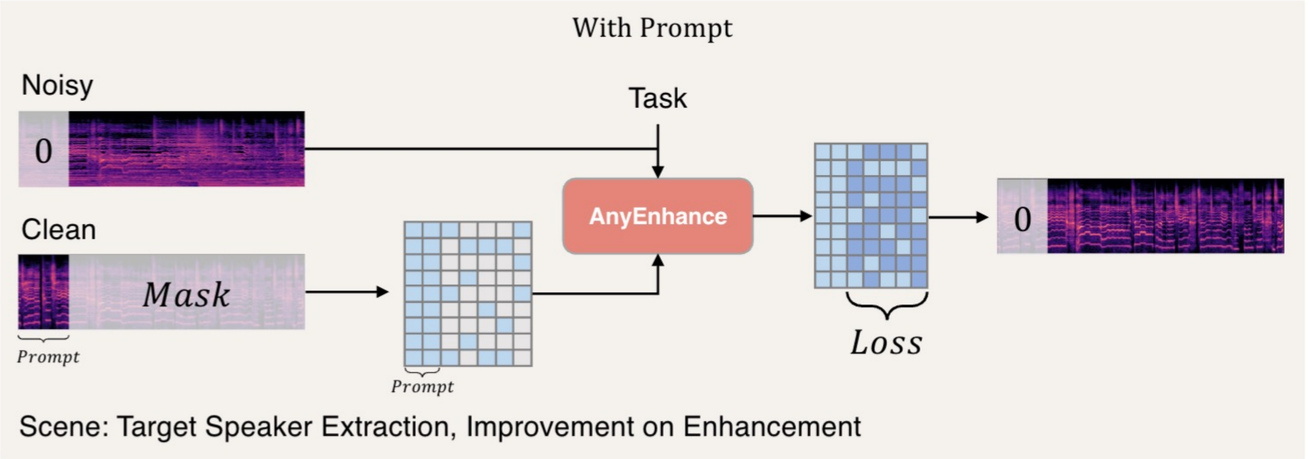
Allows the model to reference clean audio for:

- Speaker identity
- Quality preservation

Benefits:

- Enables Target Speaker Extraction
- Personalized Enhancement - further improves audio quality when speaker reference is available





Experimental Results & Demo

Model & Training:

- 363.54M parameters
- 50% keep first 3s as prompt
- 4 A800 GPUs, 5.5k token per GPU, Adam optimizer for 300k steps

Four Evaluation Groups:

- GSR Group (General Speech Restoration)
- SE (Speech Enhancement)
- SR (Super Resolution)
- TSE (Target Speaker Extraction)

TABLE II: Statistics of the speech, singing voice, noise and room impulse response (RIR) datasets used for ANYENHANCE.

Category	Datasets	Size
24kHz Speech	Emilia [14] (Filtered Subset)	20.987.49 hours
Fullband Speech	VCTK [35], HiFi-TTS [36], HQ-TTS [1], AI-Shell3 [37], Bible-TTS [38]	956.74 hours
Fullband Singing	OpenSinger [39], PopCS [40], PopBuTFy [41], Opencpop [42], M4Singer [43], SingStyle111 [44], ACESinger [45]	572.01 hours
Noise	MUSAN [46], Urgent Challenge [21], FSD50K [47], DESED [48], TUT Urban Acoustic Scenes [49], Live-performance Noise	1233.89 hours
RIR	SLR26, SLR28, Self-collected	62,668 pieces

Experimental Results & Demo

Group	Distortion(s)	Domain	Dataset	Model	SIG↑	BAK↑	OVRL↑	NISQA↑	Speech-BERTScore↑	Similarity↑
GSR	Noise, Reverb, Clipping, Bandwidth Limitation	Speech	Voicefixer GSR	TFGridNet	3.253	3.906	2.945	3.643	0.782	0.613
				NSNet2	3.011	3.969	2.758	3.433	0.728	0.615
				Voicefixer (checkpoint)	3.299	3.971	3.003	4.16	0.797	0.882
				Voicefixer (retrained)	3.3	3.984	2.996	4.054	0.818	0.884
				MaskSR	**3.408**	4.041	3.122	**4.335**	**0.832**	0.916
				ANYENHANCE	3.406	**4.073**	**3.136**	4.308	0.829	**0.924**
			Librivox GSR	TF-GridNet	3.274	3.872	2.951	3.138	0.77	0.931
				NSNet2	2.895	3.866	2.589	2.735	0.7	0.892
				Voicefixer (checkpoint)	3.365	3.971	3.085	3.77	0.706	0.864
				Voicefixer (retrained)	3.35	4.024	3.069	3.63	0.758	0.897
				MaskSR	3.499	4.133	3.258	4.155	0.779	0.94
				ANYENHANCE	**3.546**	**4.142**	**3.308**	**4.346**	**0.822**	**0.955**
		Singing	CCMusic GSR	TF-GridNet	2.764	3.37	2.362	2.396	0.57	0.81
				NSNet2	2.608	3.645	2.226	2.439	0.574	0.798
				Voicefixer (checkpoint)	2.75	3.094	2.354	2.917	0.636	0.823
				Voicefixer (retrained)	2.948	3.349	2.551	3.224	0.738	0.881
				MaskSR	3.153	3.483	2.715	3.157	0.77	0.889
				ANYENHANCE	**3.243**	**3.547**	**2.797**	**3.345**	**0.811**	**0.915**

Experimental Results & Demo

Group	Distortion(s)	Domain	Dataset	Model	SIG↑	BAK↑	OVRL↑	NISQA↑	Speech-BERTScore↑	Similarity↑
SE	Noise	Speech	DNS No Reverb*	DEMUCS	3.533	4.157	3.31	3.742	0.877	0.984
				FRCRN	3.574	4.154	3.332	4.495	**0.914**	**0.993**
				SGMSE	3.501	3.710	3.137	-	-	-
				StoRM	3.514	3.941	3.205	-	-	-
				SELM	3.508	4.096	3.258	-	-	-
				TFGridNet	3.539	4.047	3.268	4.347	0.902	0.675
				NSNet2	3.391	4.071	3.132	3.943	0.847	0.675
				Voicefixer (checkpoint)	3.504	4.109	3.253	4.274	0.819	0.956
				Voicefixer (retrained)	3.452	4.035	3.157	4.105	0.827	0.959
				MaskSR	3.616	**4.183**	3.393	4.754	0.875	0.983
				ANYENHANCE	**3.64**	4.179	**3.418**	**4.821**	0.907	0.988
	Noise, Reverb		DNS With Reverb*	DEMUCS	2.937	3.844	2.615	2.188	0.725	0.93
				FRCRN	2.933	2.923	2.279	2.27	0.783	**0.966**
				SGMSE	2.730	2.741	2.430	-	-	-
				StoRM	2.947	3.141	2.516	-	-	-
				SELM	3.160	3.577	2.695	-	-	-
				TFGridNet	3.11	3.225	2.51	2.614	**0.84**	0.686
				NSNet2	2.756	3.719	2.421	2.043	0.763	0.691
				Voicefixer (checkpoint)	3.43	4.016	3.132	**3.822**	0.711	0.91
				Voicefixer (retrained)	3.074	3.721	2.667	2.906	0.724	0.918
				MaskSR	3.396	**4.043**	3.085	3.353	0.701	0.946
				ANYENHANCE	**3.5**	4.04	**3.204**	3.722	0.738	0.951

Experimental Results & Demo

Group	Distortion(s)	Domain	Dataset	Model	SIG↑	BAK↑	OVRL↑	NISQA↑	Speech-BERTScore↑	Similarity↑
SR	Bandwidth Limitation	Speech	Voicefixer SR	Voicefixer (checkpoint)	3.405	4.029	3.11	4.131	0.873	0.882
				Voicefixer (retrained)	3.041	3.903	2.745	3.556	0.837	0.854
				AudioSR	**3.492**	4.002	3.18	4.255	0.913	0.911
				MaskSR	3.464	4.028	3.154	**4.352**	0.925	0.939
				ANYENHANCE	3.449	**4.063**	**3.156**	4.201	**0.941**	**0.943**
		Singing	CCMusic SR	Voicefixer (checkpoint)	3.179	3.534	2.743	**3.356**	0.463	0.65
				Voicefixer (retrained)	3.108	3.504	2.692	3.218	0.768	0.864
				AudioSR	3.192	3.531	2.75	2.836	0.468	0.63
				MaskSR	3.308	3.588	2.857	3.173	0.813	0.892
				ANYENHANCE	**3.339**	**3.628**	**2.899**	3.225	**0.854**	**0.919**
TSE	Other Voice	Speech	Librimix TSE*	WeSep	3.563	3.931	3.228	4.041	**0.922**	**0.991**
				TSELM	3.55	**4.084**	3.288	4.029	0.808	0.908
				ANYENHANCE	**3.638**	4.066	**3.353**	**4.277**	0.735	0.914
	Other Voice, Noise, Reverb, Clipping, Bandwidth Limitation		VCTK TSE	WeSep	2.483	2.191	1.933	1.959	0.568	0.856
				TSELM	3.345	3.875	3.004	3.388	0.58	0.81
				ANYENHANCE	**3.545**	**4.102**	**3.275**	**4.57**	**0.727**	**0.925**

Experimental Results & Demo

Ablation: Effect of Prompt Guidance

TABLE V: Effect of Prompt Guidance

Dataset	Model	SIG	BAK	OVRL	NISQA	SpeechBERTScore	Similarity
Librivox GSR	AnyEnhance	3.546	4.142	3.308	4.346	0.822	0.955
	AnyEnhance (w/ prompt)	**3.636**	**4.162**	**3.401**	**4.472**	**0.828**	**0.963**
CCMusic GSR	AnyEnhance	3.243	3.547	2.797	3.345	0.811	0.915
	AnyEnhance (w/ prompt)	**3.33**	**3.716**	**2.944**	**3.571**	**0.817**	**0.921**
Voicefixer SR	AnyEnhance	3.449	**4.063**	3.156	4.201	0.941	0.943
	AnyEnhance (w/ prompt)	**3.477**	4.041	**3.177**	**4.339**	**0.946**	**0.956**

Ablation: Joint Training on Multiple Task

TABLE VI: Effect of Any Task Training

Dataset	Model	SIG	BAK	OVRL	NISQA	SpeechBERTScore	Similarity
Librivox GSR	AnyEnhance (GSR)	3.438	4.112	3.179	3.827	0.74	**0.913**
	AnyEnhance	**3.504**	**4.135**	**3.255**	**4.191**	**0.762**	0.91
DNS No Reverb*	AnyEnhance (SE)	**3.646**	4.189	3.425	4.765	**0.877**	**0.981**
	AnyEnhance	3.643	**4.205**	**3.432**	**4.774**	0.876	0.98
Voicefixer SR	AnyEnhance (SR)	3.443	4.04	3.141	4.144	**0.943**	**0.941**
	AnyEnhance	**3.476**	**4.085**	**3.197**	**4.339**	0.923	0.926
VCTK TSE	AnyEnhance (TSE)	3.479	4.004	3.17	4.251	0.596	0.888
	AnyEnhance	**3.465**	**4.077**	**3.185**	**4.451**	**0.653**	**0.9**

Experimental Results & Demo

Ablation: Data Simulation & Singing Voice Data

$$d_{\text{noise}} \circ d_{\text{bw}} \circ d_{\text{clip}} \circ d_{\text{reverb}}$$

$$D(x) = d_{\text{bw}} \circ d_{\text{clip}} \circ d_{\text{reverb}} \circ d_{\text{noise}} \circ d_{\text{vocal\_effect}}$$

Where $d_{\text{vocal\_effect}} = d_{\text{reverb}} \circ d_{\text{eq}}$ is the vocal effect chain.

Dataset	Model	SIG↑	BAK↑	OVRL↑	NISQA↑	SpeechBERTScore↑	Similarity↑
Librivox GSR	ANYENHANCE (w/o improved data simulation)	3.361	4.096	3.093	3.766	0.731	0.888
	ANYENHANCE (w/ improved data simulation)	**3.504**	**4.135**	**3.255**	**4.191**	**0.762**	**0.91**
Voicefixer GSR	ANYENHANCE (w/o improved data simulation)	3.355	**4.033**	3.066	4.145	0.798	0.897
	ANYENHANCE (w/ improved data simulation)	**3.406**	4.024	**3.11**	**4.339**	**0.832**	**0.909**

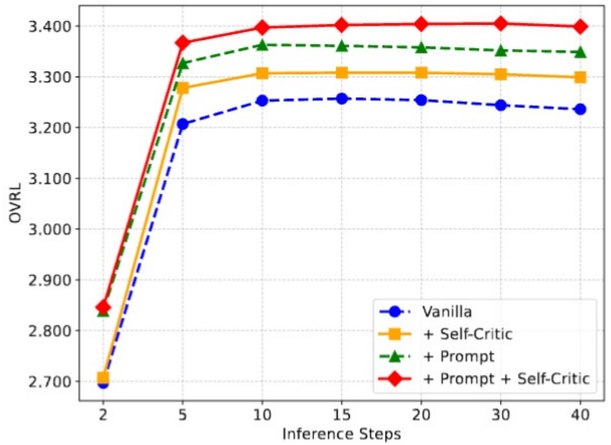
Dataset	+ Improved Simulation	+ Singing Voice Data	SIG↑	BAK↑	OVRL↑	NISQA↑	SpeechBERTScore↑	Similarity↑
CCMusic GSR	✗	✗	2.981	3.736	2.542	2.916	0.559	0.776
	✓	✗	3.17	**3.77**	2.75	**3.406**	0.598	0.839
	✗	✓	3.144	3.537	2.694	3.124	0.739	0.863
	✓	✓	**3.213**	3.647	**2.809**	3.358	**0.759**	**0.889**

Experimental Results & Demo

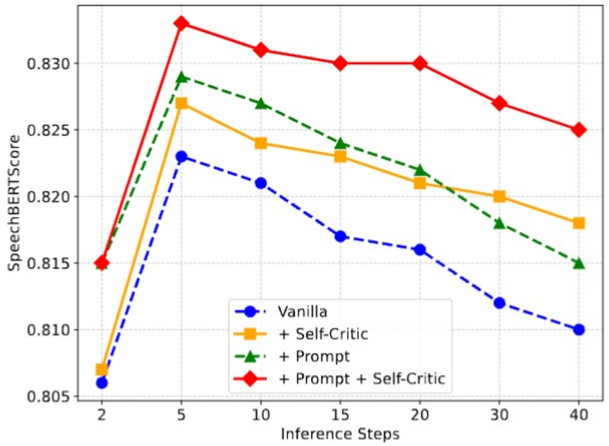
Ablation:

Self-Critic Effect

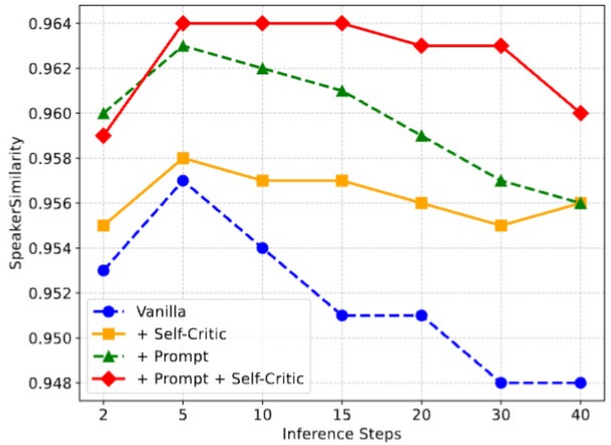
+ Prompt Guidance



(a) OVRL



(b) SpeechBERTScore



(c) Speaker Similarity

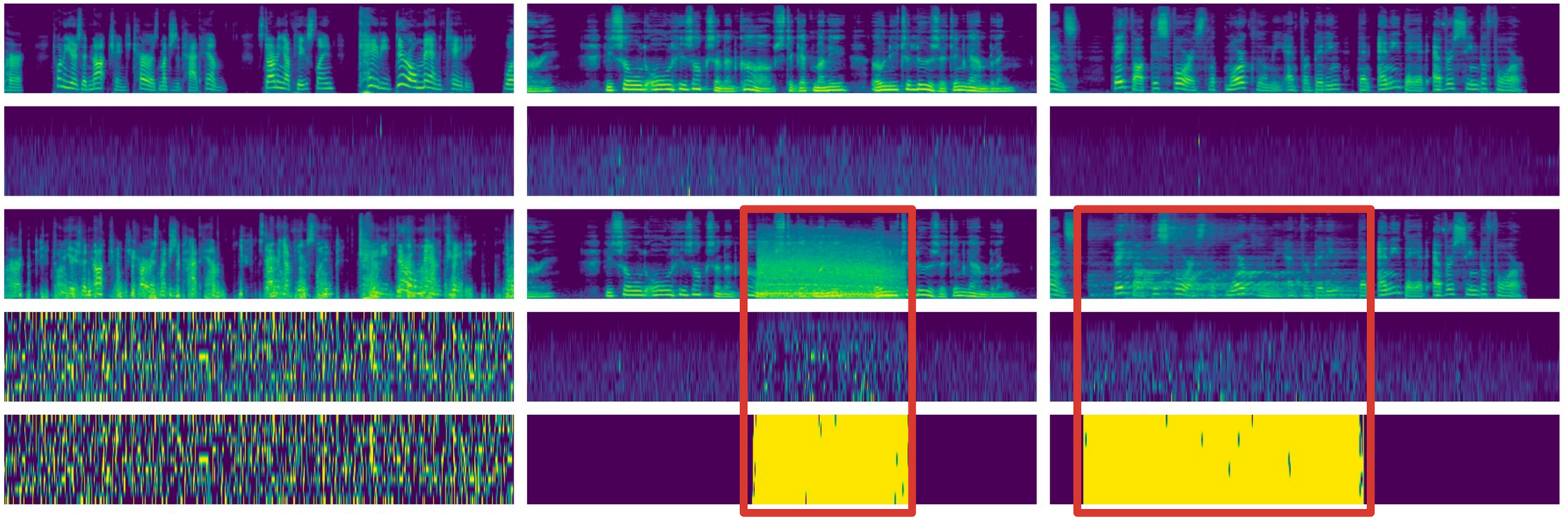
Original Audio

Critic Output

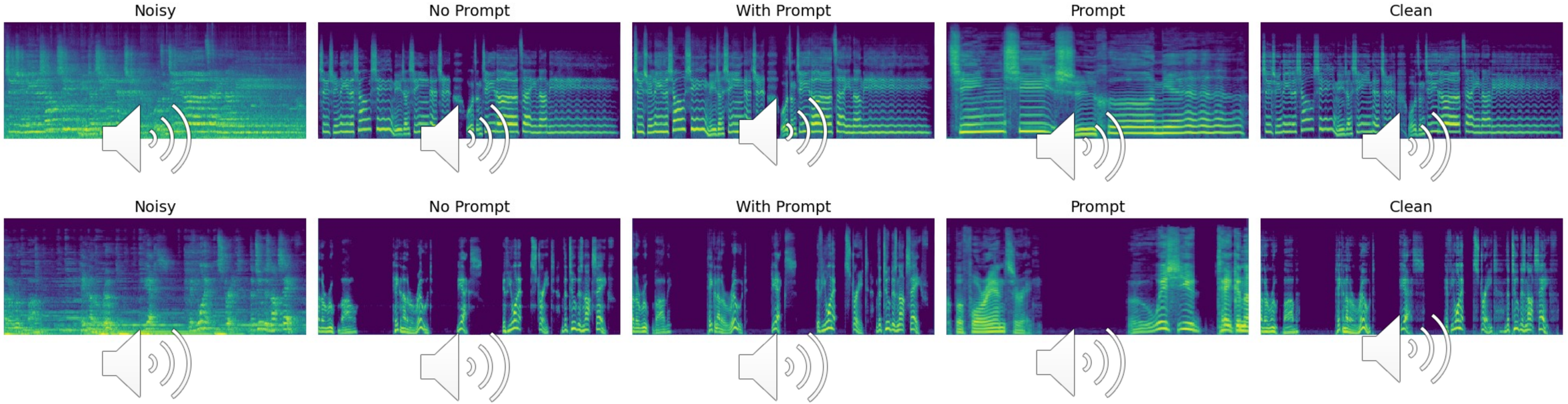
Distorted Audio

Critic Output

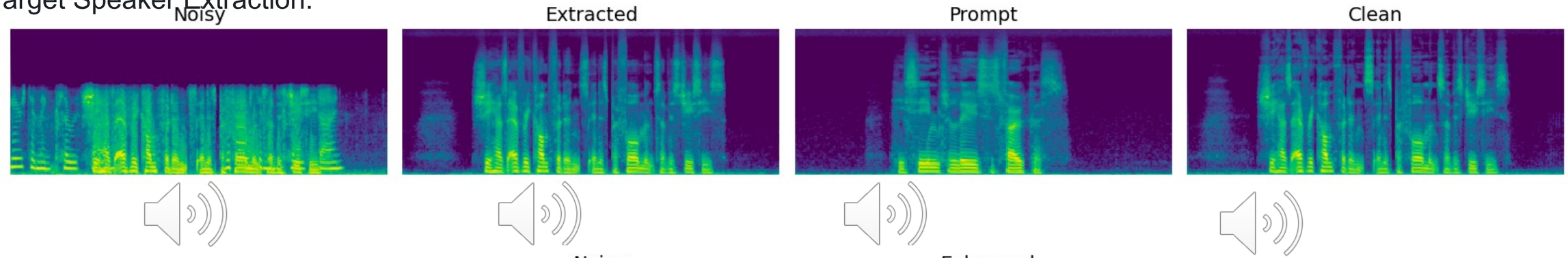
Token Diff



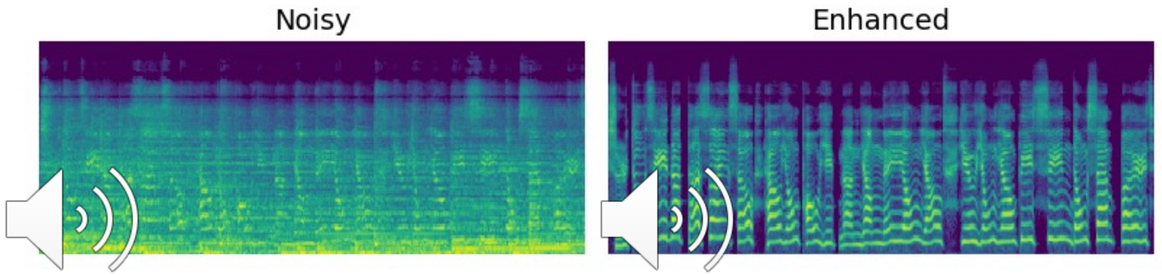
Experimental Results & Demo



Target Speaker Extraction:



Real-World Live Performance:



Thanks!

- **Paperlink: https://arxiv.org/abs/2501.15417**
- **Demopage: https://amphionspace.github.io/anyenhance**

collaboration welcome!

interest: audio enhancement/music generation