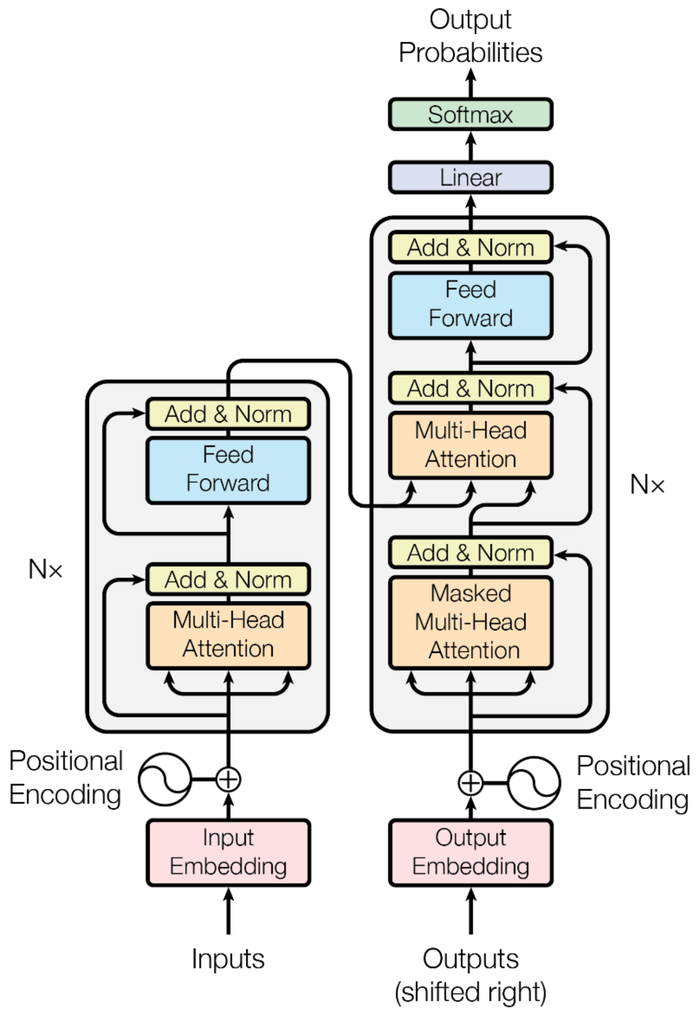
Voice Enhancement in the Era of Large Language Models

Junan Zhang
CUHKSZ
junanzhang@link.cuhk.edu.cn

- **https://viewfinder-annn.github.io/**
- **~~first~~ second-year phd student, advised by prof. zhizheng wu, bachelor in fudan university**
- **interest: audio/music enhancement & generation**
- **collaboration welcome!**

Background

Era of Large Language Models



Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Sutskever [**1]

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou
Google Research, Brain Team
{jasonwei,dennyzhou}@google.com

Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*
Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray
John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens
Amanda Askell† Peter Welinder Paul Christiano*†
Jan Leike* Ryan Lowe*
OpenAI

Background for

Voice Enhancement in the Era of Large Language Models:

1. **AnyEnhance:** Multitask Enhancement Model with Prompt Guidance
2. **GenSR-Pref:** Preference Alignment for Generative Speech Restoration

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Sutskever [**1]

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou

Google Research, Brain Team
{jasonwei,dennyzhou}@google.com

Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*
Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray
John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens
Amanda Askell† Peter Welinder Paul Christiano*†
Jan Leike* Ryan Lowe*

OpenAI

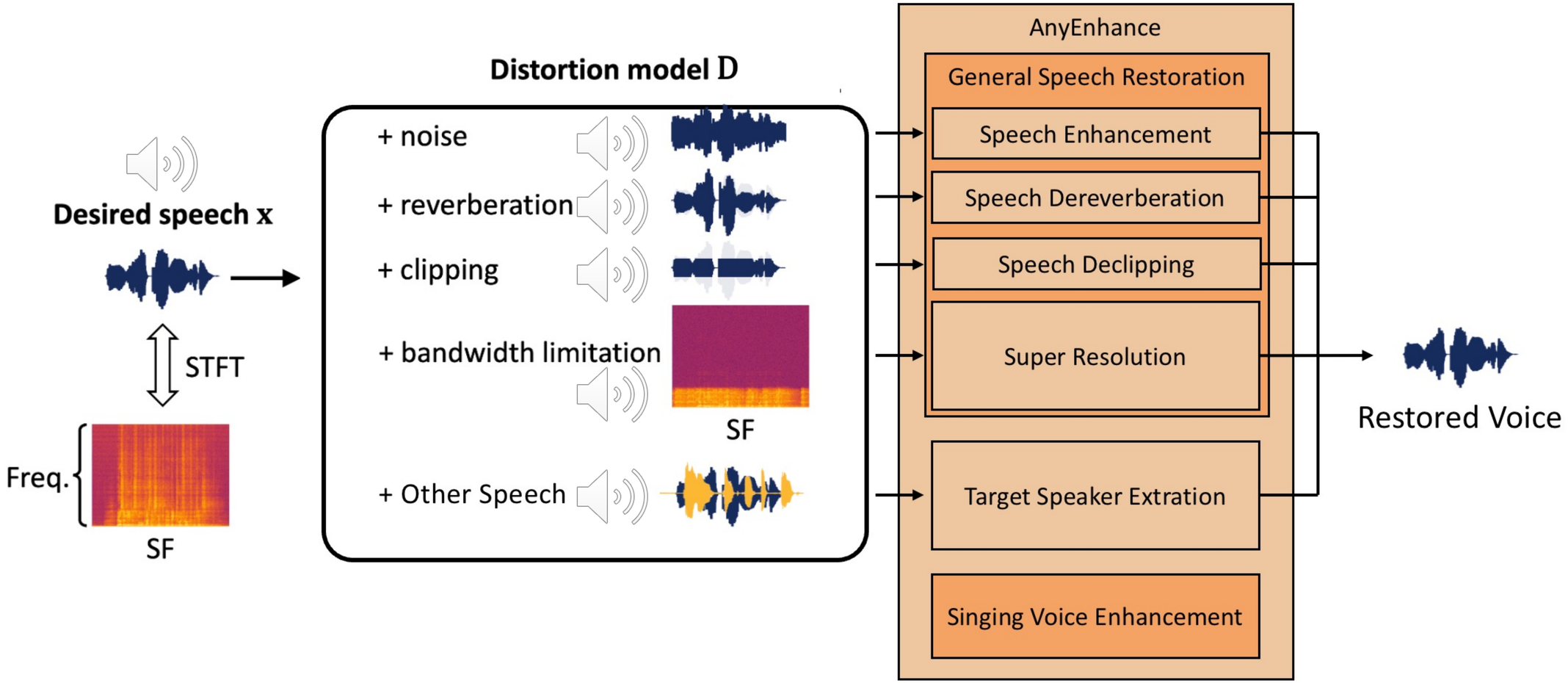
AnyEnhance: Multitask Model with Prompt Guidance

ANYENHANCE: A Unified Generative Model With Prompt-Guidance and Self-Critic for Voice Enhancement

Junan Zhang, Jing Yang, Zihao Fang, Yuancheng Wang, Zehua Zhang, Zhuo Wang, Fan Fan, and Zhizheng Wu, *Senior Member, IEEE*

- **TASLP 2025**
- **Paperlink: https://arxiv.org/abs/2501.15417**
- **Demopage: https://amphionspace.github.io/anyenhance**

AnyEnhance: Multitask Model with Prompt Guidance



Enhancement Model in the Era of Large Language Models:

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

AnyEnhance: Multitask Model with Prompt Guidance

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:

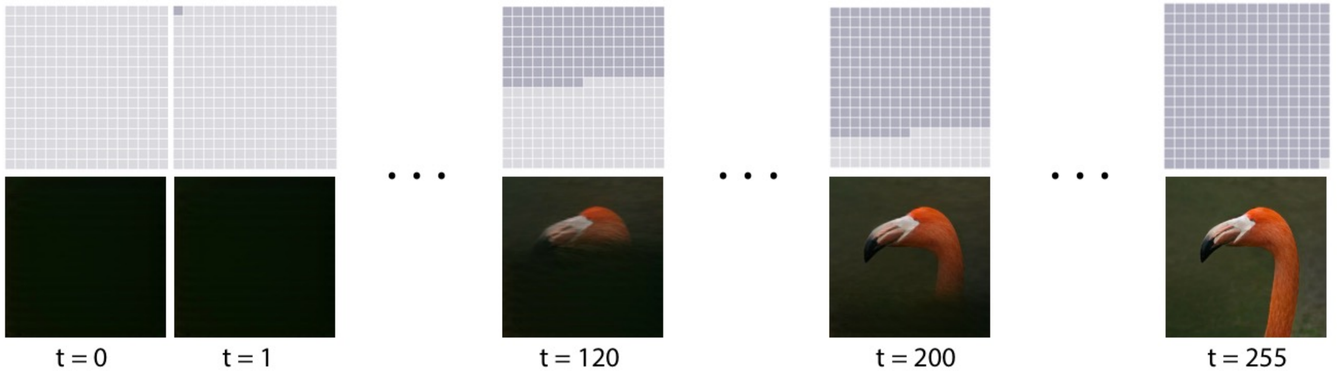
1. Multi-Task, Multi-Domain, Simultaneous
2. Supports prompt audio

What's more:

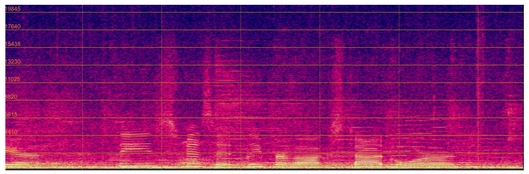
3. Self-critic for more stable sampling

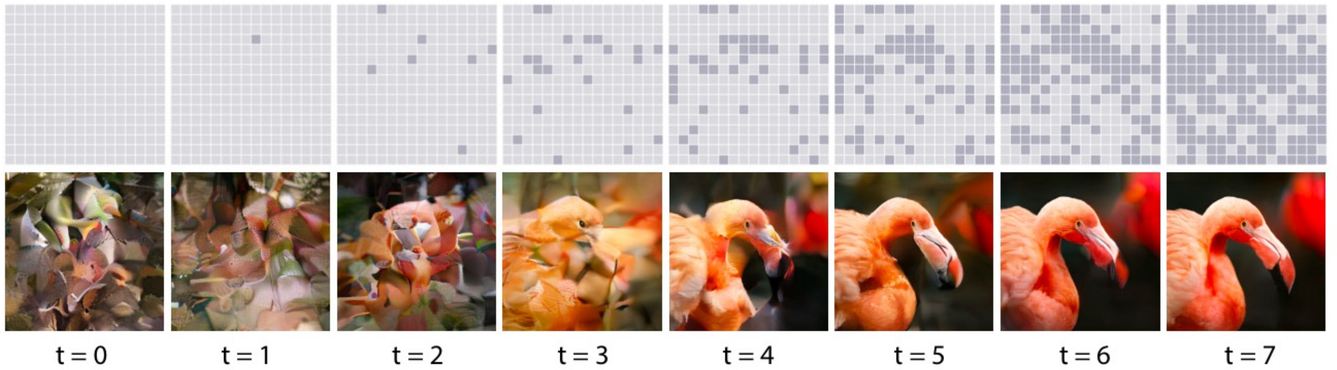
AnyEnhance: Multitask Model with Prompt Guidance

masked generative models

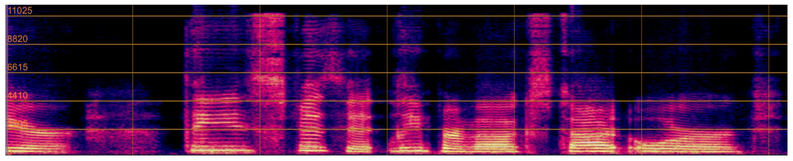


input

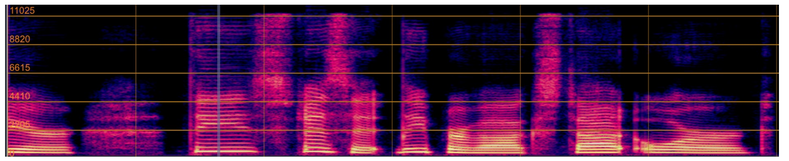




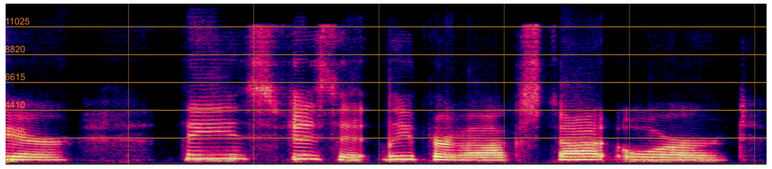
t=5



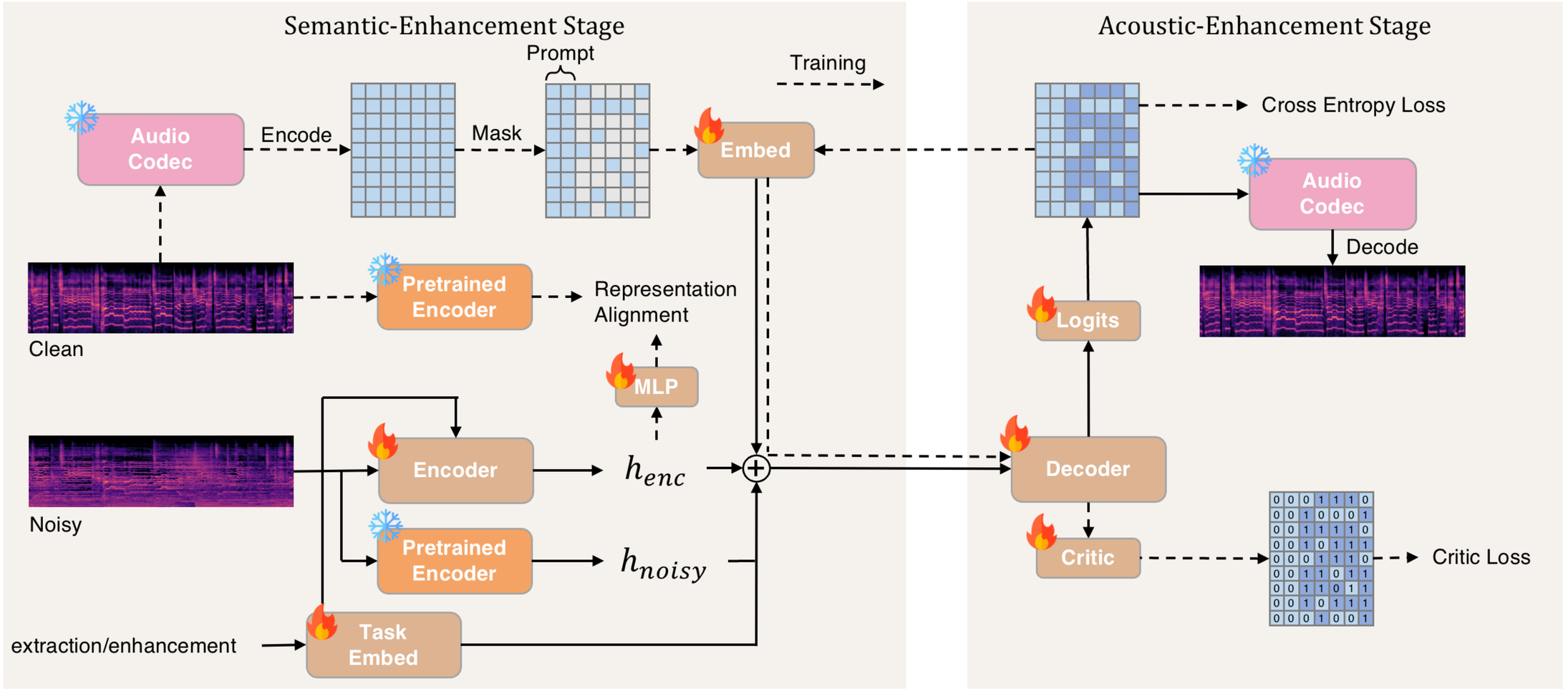
t=50



t=100



AnyEnhance: Multitask Model with Prompt Guidance



2 stages:

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

AnyEnhance: Multitask Model with Prompt Guidance

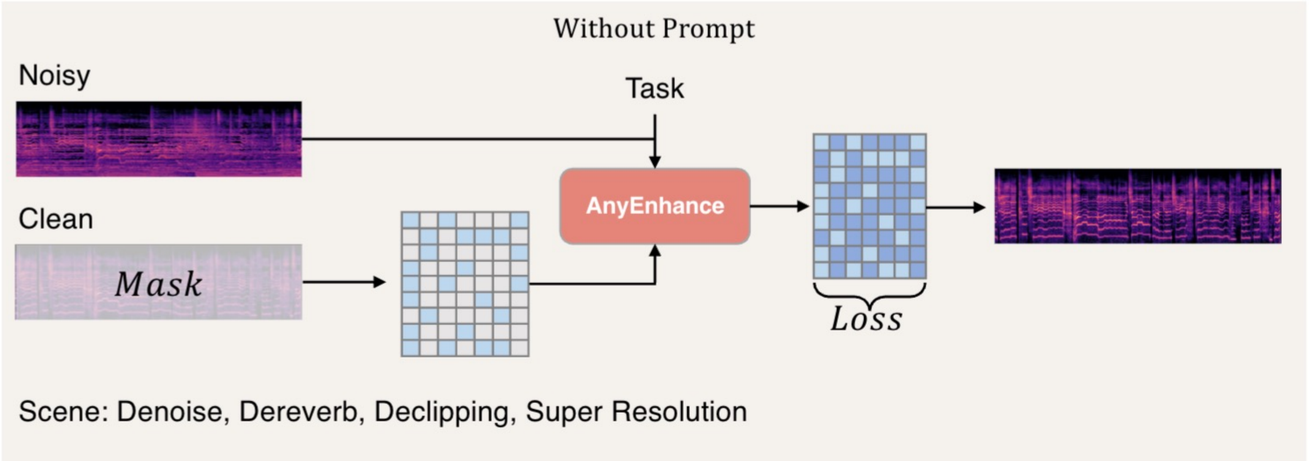
Inspired by zero-shot tts, during training, we preserve a portion of clean audio with a certain probability.

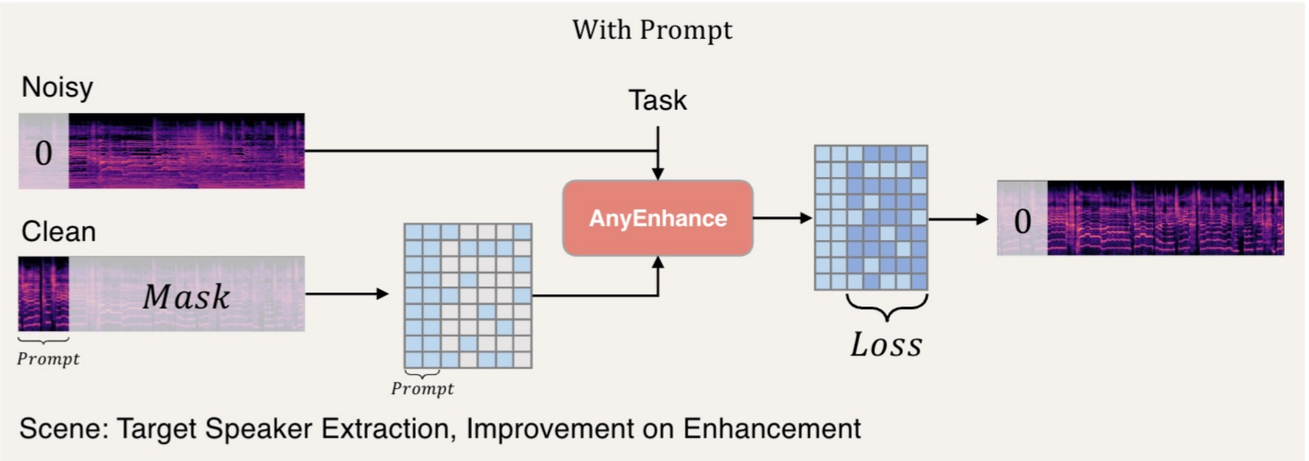
Allows the model to reference clean audio for:

- Speaker identity
- Quality preservation

Benefits:

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





AnyEnhance: Multitask Model with Prompt Guidance

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

AnyEnhance: Multitask Model with Prompt Guidance

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**

AnyEnhance: Multitask Model with Prompt Guidance

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

AnyEnhance: Multitask Model with Prompt Guidance

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**

AnyEnhance: Multitask Model with Prompt Guidance

Ablation: Effect of Prompt Guidance

TABLE V: Effect of Prompt Guidance

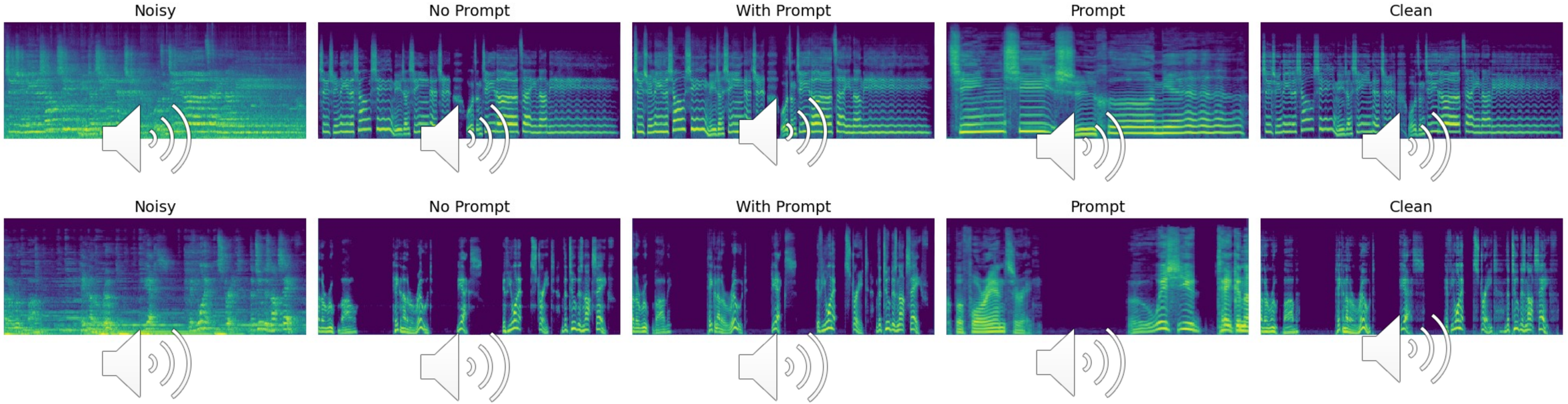
Dataset	Model	SIG	BAK	OVRL	NISQA	SpeechBERTScore	Similarity
Librivox GSR	AnyEnhance	3.546	4.142	3.308	4.346	0.822	0.955
	AnyEnhance (w/ prompt)	**3.636**	**4.162**	**3.401**	**4.472**	**0.828**	**0.963**
CCMusic GSR	AnyEnhance	3.243	3.547	2.797	3.345	0.811	0.915
	AnyEnhance (w/ prompt)	**3.33**	**3.716**	**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

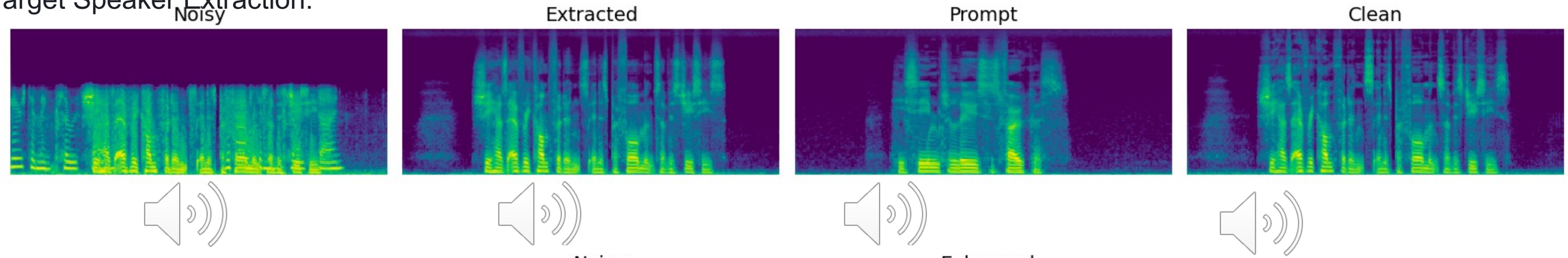
TABLE VI: Effect of Any Task Training

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**

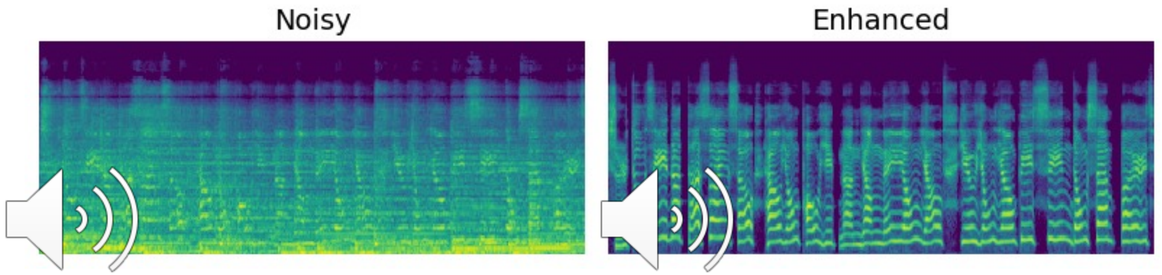
AnyEnhance: Multitask Model with Prompt Guidance



Target Speaker Extraction:



Real-World Live Performance:



AnyEnhance: Multitask Model with Prompt Guidance

What's next step once we have a powerful *pre-trained* model?

Training language models to follow instructions with human feedback

Long Ouyang* **Jeff Wu*** **Xu Jiang*** **Diogo Almeida*** **Carroll L. Wainwright***

Pamela Mishkin* **Chong Zhang** **Sandhini Agarwal** **Katarina Slama** **Alex Ray**

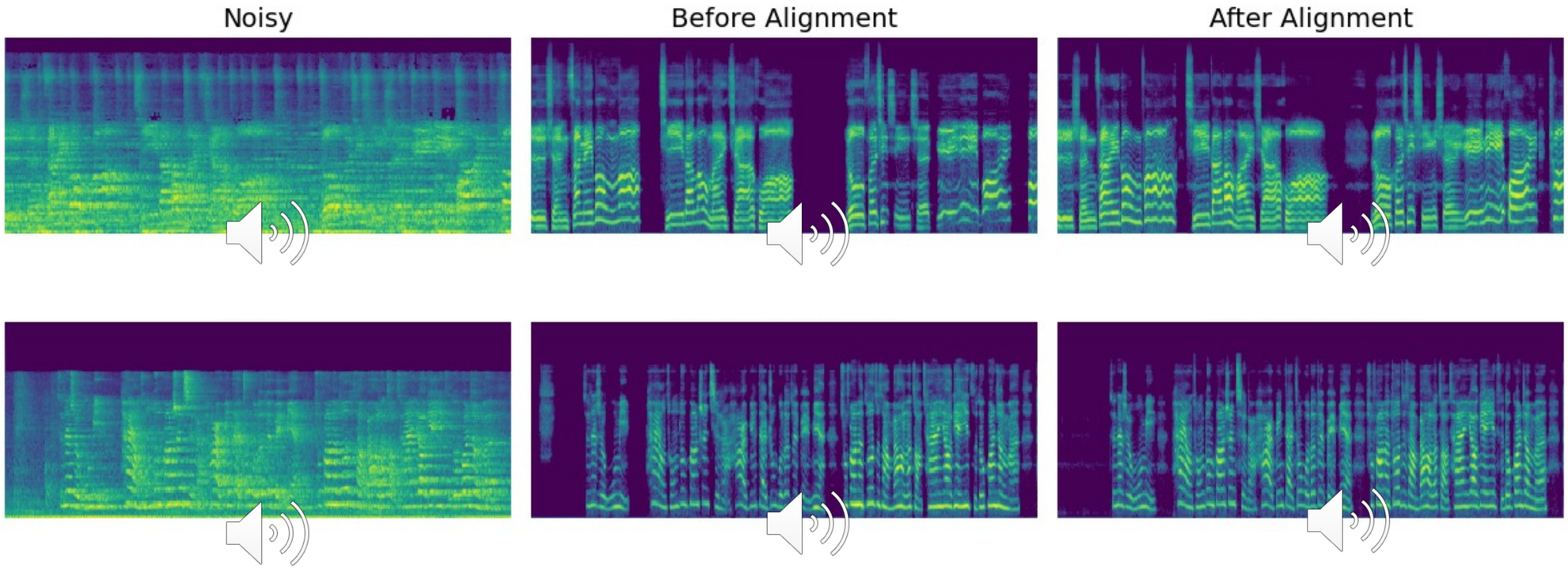
John Schulman **Jacob Hilton** **Fraser Kelton** **Luke Miller** **Maddie Simens**

Amanda Askell† **Peter Welinder** **Paul Christiano*†**

Jan Leike* **Ryan Lowe***

OpenAI

AnyEnhance Can Be Better!



GenSR-Pref: Preference Alignment for Enhancement

Multi-Metric Preference Alignment for Generative Speech Restoration

Junan Zhang[1], Xueyao Zhang[1], Jing Yang[2], Yuancheng Wang[1], Fan Fan[2], Zhizheng Wu[1]*

[1]The Chinese University of Hong Kong, Shenzhen
[2]Central Media Technology Institute, Huawei
junanzhang@link.cuhk.edu.cn, wuzhizheng@cuhk.edu.cn

- **AAAI 2026 (Phase 2)**
- **Paperlink: https://arxiv.org/pdf/2508.17229**
- **Demopage: https://gensr-pref.github.io/**

GenSR-Pref: Preference Alignment for Enhancement

The Good:

- Advances in Generative models have achieved state-of-the-art results in speech restoration tasks.
- They can generate high-fidelity audio even from severely degraded inputs.

The Bad:

- Misalignment Training objectives (e.g., likelihood maximization) often misalign with human perceptual preferences.

A Promissing Direction:

- Post-training alignment (e.g., DPO, RLHF) has been highly successful in other domains like NLP and Vision.
- Can we effectively apply this to speech restoration?

GenSR-Pref: Preference Alignment for Enhancement

Key Challenges for Alignment in Speech Restoration:

1. Defining a Faithful Preference Signal:
- How can we create an automated proxy for complex human hearing, which values clarity, naturalness, and fidelity?
- A naive signal can lead to **reward hacking**: improving one metric while degrading others.

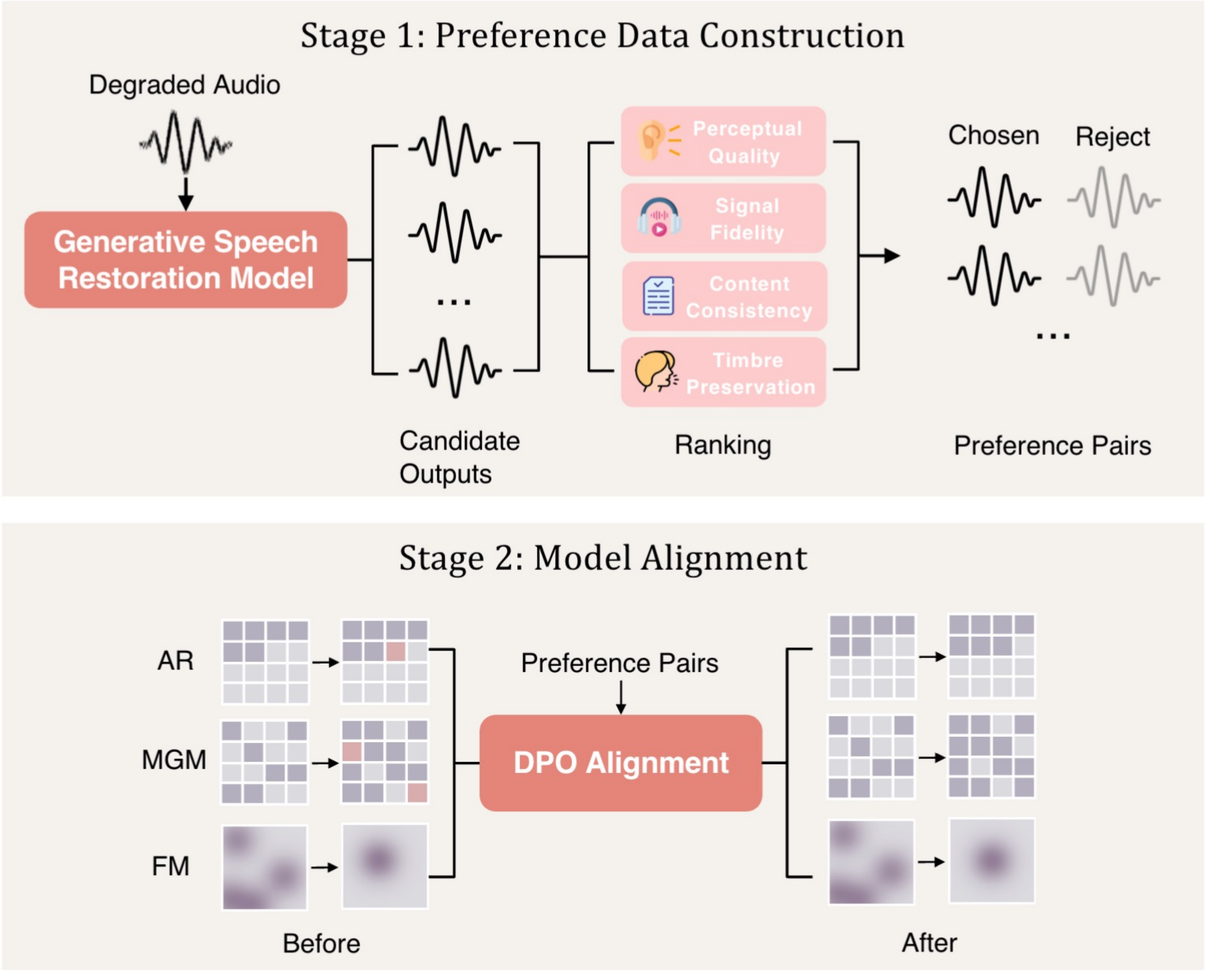
2. Curating High-Quality Preference Data:
- Given a preference signal, what is an effective strategy to build a dataset that robustly guides the model?

3. Ensuring Holistic Improvement:
- How do we ensure the model learns genuine, overall quality improvement rather than exploiting metric biases?

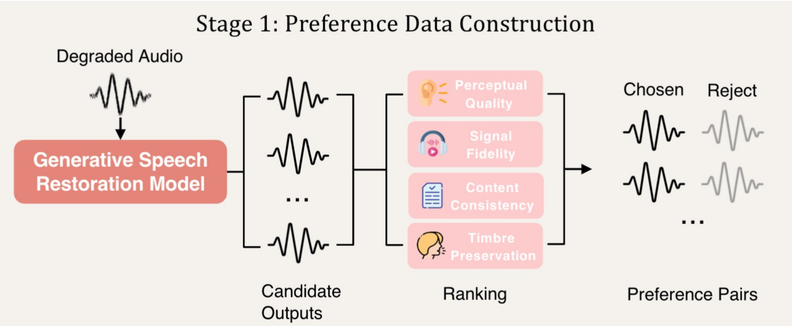
GenSR-Pref: Preference Alignment for Enhancement

Our Proposal: A Multi-Metric Preference Alignment Strategy

- We propose a two-stage, model-agnostic strategy to address these challenges.



GenSR-Pref: Preference Alignment for Enhancement



The "Unanimous Agreement" Criterion

A preference pair (y_w, y_l) is created only if the winning sample y_w scores higher than the losing sample y_l on **all four** of the following complementary metrics:

- **Perceptual Quality** (NISQA): Overall naturalness and listening comfort.
- **Signal Fidelity** (DNSMOS): Low distortion and background noise.
- **Content Consistency** (SpeechBERTScore): Preserves linguistic content.
- **Timbre Preservation** (Speaker Similarity): Maintains speaker identity.

GenSR-Pref: Preference Alignment for Enhancement

Direct Preference Optimization (DPO)

DPO is a simple and stable method to align models with preferences, avoiding the complexities of reinforcement learning.
The core idea is to directly optimize the policy using a contrastive loss on preferred (y_w) and dispreferred (y_l) pairs via Bradley-Terry model:
$P(y_w \succ y_l|x) = \sigma(r(x, y_w) - r(x, y_l))$.

Paradigm	RL Objective (with KL Regularization)	Implicit Reward Mapping	Final DPO Loss Formulation
AR	$\max_{p_\theta} \mathbb{E}_{y\sim p_\theta(y\|x)}[r(x,y)] - \beta D_{\mathrm{KL}}\left[p_\theta(y\|x) \,\|\, p_{\mathrm{ref}}(y\|x)\right]$	$r(x,y) = \beta \log \frac{p^*_\theta(y\|x)}{p_{\mathrm{ref}}(y\|x)} + \beta \log Z(x)$	$\mathcal{L}_{\mathrm{DPO}} = -\mathbb{E}_{\mathcal{D}}\left[\log \sigma\left(\beta \log \frac{p_\theta(y_w\|x)}{p_{\mathrm{ref}}(y_w\|x)} - \beta \log \frac{p_\theta(y_l\|x)}{p_{\mathrm{ref}}(y_l\|x)}\right)\right]$
MGM	$\max_{p_\theta} \mathbb{E}_{y_0,t,x}[r(y_0,x)] - \beta D_{\mathrm{KL}}\left[p_\theta(y_0\|y_t,x) \,\|\, p_{\mathrm{ref}}(y_0\|y_t,x)\right]$	$r(y_0,x) = \beta \log \frac{p^*_\theta(y_0\|y_t,x)}{p_{\mathrm{ref}}(y_0\|y_t,x)} + \beta \log Z(y_t,x)$	$\mathcal{L}_{\text{DPO-MGM}} = -\mathbb{E}_{\mathcal{D},t}\left[\log \sigma\left(\beta \log \frac{p_\theta(y_0^w\|y_t^w,x)}{p_{\mathrm{ref}}(y_0^w\|y_t^w,x)} - \beta \log \frac{p_\theta(y_0^l\|y_t^l,x)}{p_{\mathrm{ref}}(y_0^l\|y_t^l,x)}\right)\right]$
FM	$\max_{p_\theta} \mathbb{E}_{y_1,t,x}[r(y_1,x)] - \beta D_{\mathrm{KL}}\left[p_\theta(y_1\|y_t,t,x) \,\|\, p_{\mathrm{ref}}(y_1\|y_t,t,x)\right]$	$r(y_1,x) = \beta \log \frac{p^*_\theta(y_1\|y_t,t,x)}{p_{\mathrm{ref}}(y_1\|y_t,t,x)} + \beta \log Z(y_t,t,x)$	$\mathcal{L}_{\text{DPO-FM}} = -\mathbb{E}_{\mathcal{D},t} \log \sigma\left(-\beta(\Delta_w - \Delta_l)\right)$, where $\Delta = \|v_\theta(y_t,t,x) - (y_1 - y_0)\|_2^2 - \|v_{\mathrm{ref}}(y_t,t,x) - (y_1 - y_0)\|_2^2$

GenSR-Pref: Preference Alignment for Enhancement

Models

We evaluate our strategy across three diverse generative paradigms:

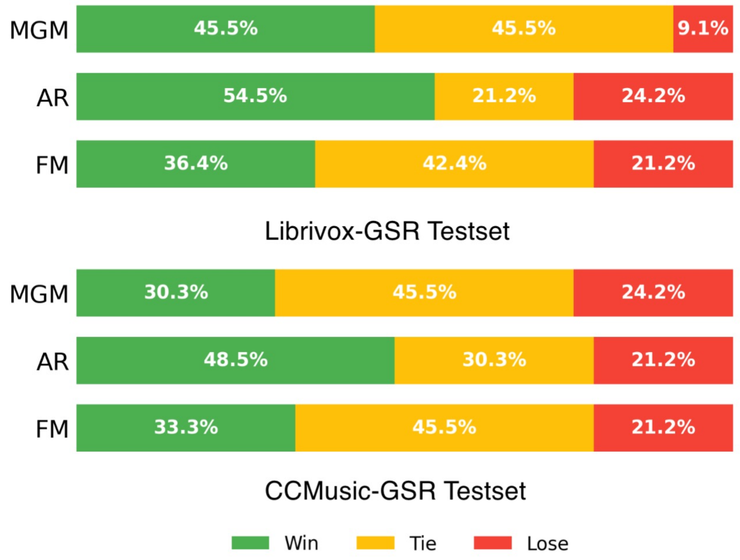
- **MGM**: `AnyEnhance` (pre-trained)
- **AR**: `AR+Soundstorm` (trained by us)
- **FM**: `Flow-SR` (trained by us)

Evaluation

- **Benchmarks**: General Speech Restoration (Voicefixer-GSR, LibriVox-GSR) and Speech Enhancement (DNS benchmark).
- **Metrics**: DNSMOS (SIG, BAK, OVRL), NISQA, SpeechBERTScore, Speaker Similarity.
- **Subjective Tests**: A/B preference tests with human listeners.

GenSR-Pref: Preference Alignment for Enhancement

Dataset	Model	Type	DPO-aligned?	SIG↑	BAK↑	OVRL↑	NISQA↑	Speech-BERTScore↑	Similarity↑
Voicefixer-GSR	Voicefixer	DISC.	-	3.299	3.971	3.003	4.160	0.797	0.882
	MaskSR	MGM	-	3.445	3.971	3.128	-	-	-
	AnyEnhance	MGM	✗	3.406	4.073	3.136	4.308	0.829	0.924
			✓	**3.532**	**4.091**	**3.267**	**4.639**	**0.834**	**0.935**
	AR+Soundstorm	AR	✗	3.550	4.097	3.294	4.556	0.788	0.894
			✓	**3.564**	**4.144**	**3.331**	**4.850**	**0.803**	**0.904**
	Flow-SR	FM	✗	3.398	3.969	3.104	4.010	0.812	0.918
			✓	**3.483**	**4.092**	**3.230**	**4.672**	**0.830**	**0.924**
Librivox-GSR	AnyEnhance	MGM	✗	3.546	4.142	3.308	4.346	0.822	**0.955**
			✓	**3.690**	**4.201**	**3.475**	**4.865**	**0.828**	0.954
	AR+Soundstorm	AR	✗	3.663	4.134	3.419	4.535	0.783	0.922
			✓	**3.693**	**4.193**	**3.478**	**4.923**	**0.793**	**0.924**
	Flow-SR	FM	✗	3.550	4.062	3.281	4.184	0.791	**0.931**
			✓	**3.602**	**4.152**	**3.368**	**4.825**	**0.801**	0.930
CCMusic-GSR	AnyEnhance	MGM	✗	3.243	3.547	2.797	3.345	0.811	**0.915**
			✓	**3.440**	**3.827**	**3.062**	**4.154**	**0.817**	0.909
	AR+Soundstorm	AR	✗	3.378	3.693	2.956	3.948	0.710	**0.854**
			✓	**3.460**	**3.865**	**3.094**	**4.438**	**0.713**	0.853
	Flow-SR	FM	✗	3.298	3.540	2.813	3.897	0.733	**0.883**
			✓	**3.378**	**3.770**	**2.971**	**4.371**	**0.740**	0.881



GenSR-Pref: Preference Alignment for Enhancement

Ablation: Multi-Metric vs Single-Metric

AR

Criterion	SIG	BAK	OVRL	NISQA	SBERT	SIM
-	3.550	4.097	3.294	4.556	0.788	0.894
Multi-Metric	**3.564**	**4.144**	**3.331**	**4.850**	0.803	**0.904**
NISQA	3.531	4.137	3.300	4.810	0.785	0.896
OVRL	3.561	4.117	3.317	4.600	0.792	0.896
SIM	3.537	4.101	3.285	4.577	0.792	0.901
SBERT	3.540	4.109	3.291	4.612	**0.804**	0.901

MGM

Criterion	SIG	BAK	OVRL	NISQA	SBERT	SIM
-	3.406	4.073	3.136	4.308	0.829	0.924
Multi-Metric	3.532	4.091	3.267	4.639	0.834	0.935
NISQA	3.429	4.096	3.177	4.715	0.811	0.930
OVRL	3.536	4.088	3.274	4.459	0.820	0.928
SBERT	3.407	4.031	3.121	4.364	0.861	0.929
SIM	3.372	4.060	3.107	4.381	0.841	0.936

FM

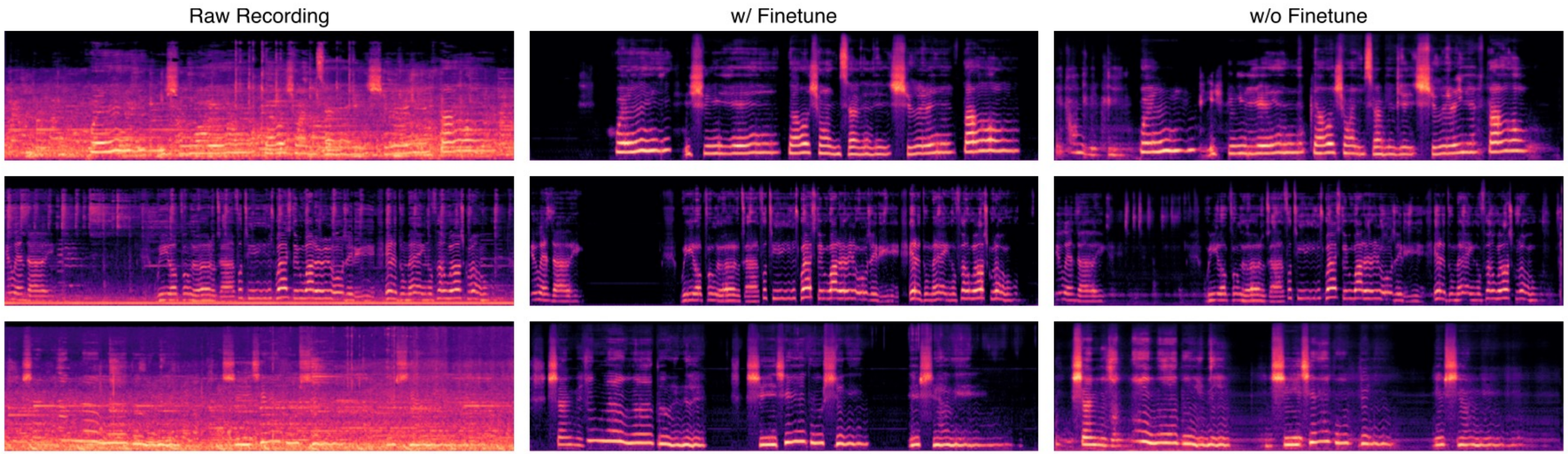
Criterion	SIG	BAK	OVRL	NISQA	SBERT	SIM
-	3.398	3.969	3.104	4.010	0.812	0.918
Multi-Metric	3.483	4.092	3.230	4.672	0.830	0.924
NISQA	3.415	4.035	3.147	4.495	0.814	0.920
OVRL	3.464	4.010	3.179	4.206	0.815	0.918
SBERT	3.413	4.002	3.129	4.209	0.826	0.919
SIM	3.394	3.951	3.091	4.095	0.810	0.921

GenSR-Pref: Preference Alignment for Enhancement

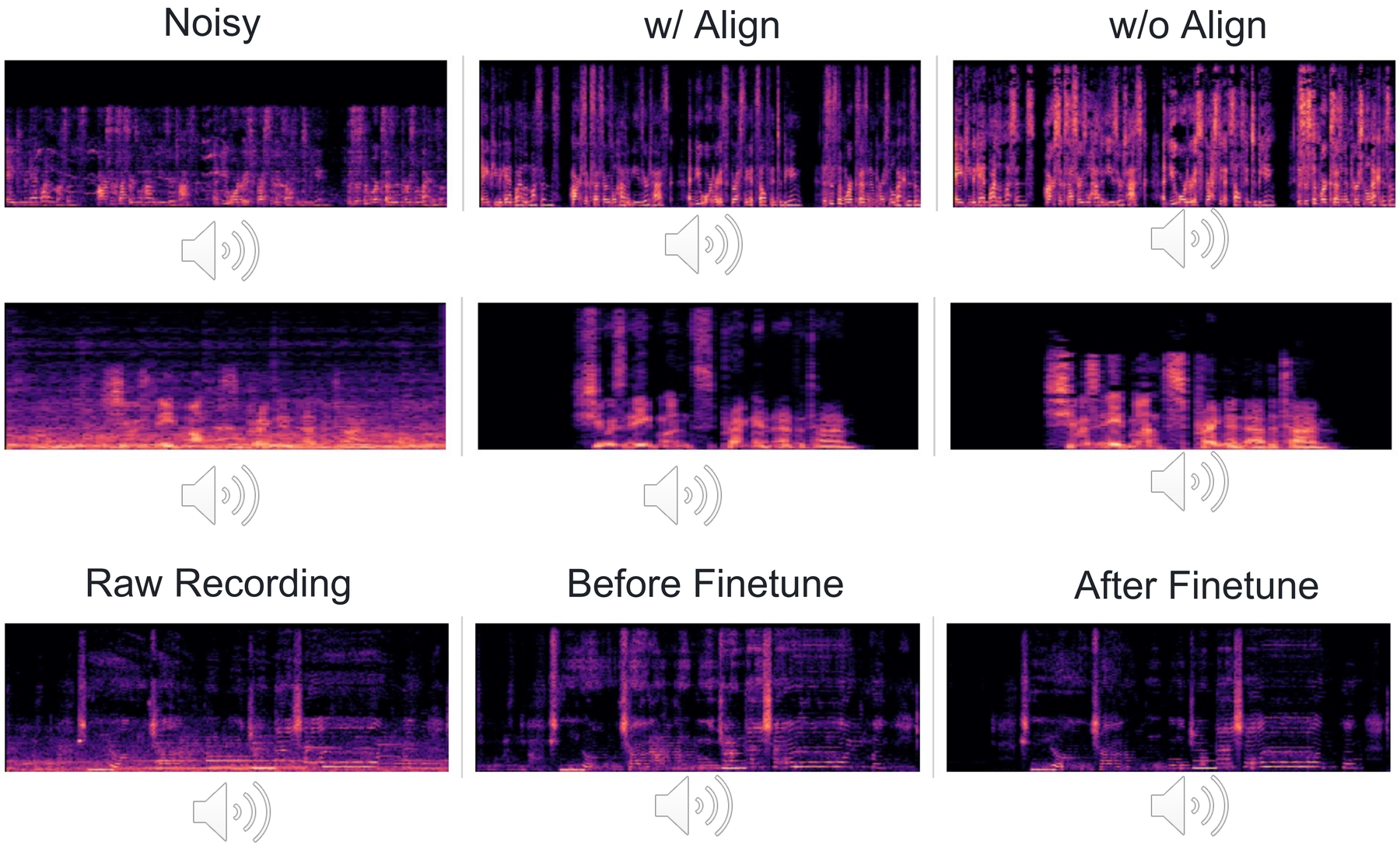
Application: Empowering Discriminative Models via Pseudo-Labeling

Our aligned models can act as powerful ``data annotators", generating high-quality pseudo-labels to train discriminative models in data-scarce scenarios like singing voice restoration

Model	SIG	BAK	OVRL	NISQA
Voicefixer (before)	2.657	3.080	2.295	2.919
Voicefixer (after)	**3.096**	**3.745**	**2.756**	**3.312**



GenSR-Pref: Preference Alignment for Enhancement



Summary

Voice Enhancement in the Era of Large Language Models:

1. **AnyEnhance:** Multitask Enhancement Model with Prompt Guidance

   Generative models can effectively handle multiple degradations, as well as zero-shot prompt guidance.

2. **GenSR-Pref:** Preference Alignment for Generative Speech Restoration

   Post-training alignment is feasible on voice enhancement tasks.

Future Work:

- **Human-based preference alignment / specialized preference alignment.**
- **Models with a broader foundation: expanded domain coverage (e.g. music), task generalization.**
- **Adaptation of generative foundation models: peft / posterior sampling / inversion.**

Thanks!

- **AnyEnhance: https://arxiv.org/abs/2501.15417**
- **GenSR-Pref: https://arxiv.org/abs/2508.17229**