Audio-to-Audio Schrödinger Bridges

Zhifeng Kong* NVIDIA
Kevin J Shih* NVIDIA
Weili Nie NVIDIA
Arash Vahdat NVIDIA
Sang-gil Lee NVIDIA
João Felipe Santos NVIDIA
Ante Jukić NVIDIA
Rafael Valle NVIDIA
Bryan Catanzaro NVIDIA

Abstract

Real-world audio is often degraded by numerous factors. This work presents an audio restoration model tailored for high-res (44.1kHz) music. Our model, Audio-to-Audio Schrödinger Bridges (A^2SB), is capable of both bandwidth extension (predicting high-frequency components) and inpainting (re-generating missing segments). Critically, it is end-to-end – requiring no vocoder to predict waveform outputs, able to restore hour-long audio inputs, and trained on permissively licensed music data. A^2SB is capable of achieving state-of-the-art bandwidth extension and inpainting quality on several out-of-distribution music test sets. Code and model: `https://github.com/NVIDIA/diffusion-audio-restoration`.

1 Introduction

Real world audio is subject to numerous degradation factors such as recording devices, data compression, and online transfers – resulting in low sampling rate and content loss. These problems are usually ill-posed [Narayanaswamy et al., 2021, Moliner et al., 2023] and solved with data-driven generative models. Bandwidth extension methods have been proposed to up-sample the audio [Lee and Han, 2021, Liu et al., 2022, Serrà et al., 2022, Moliner and Välimäki, 2022, Shuai et al., 2023, Yu et al., 2023, Kim et al., 2024, Liu et al., 2024, Ku et al., 2024, Yun et al., 2025], and inpainting methods have been developed to predict segments where audio is missing [Marafioti et al., 2019, 2020, Borsos et al., 2022, Liu et al., 2023b, Moliner and Välimäki, 2023, Asaad et al., 2024]. Many of these methods are task-specific, designed for the speech domain, or trained to only restore the degraded magnitude – which requires an additional vocoder to transform the restored magnitude into a waveform. Our work investigates high-res (44.1kHz) music restoration, a more challenging task than speech restoration in terms of typical bandwidth. We aim to tackle bandwidth extension and inpainting in a single model and also to build an end-to-end trainable generative model for audio restoration without using a separate vocoder or a codec. To achieve this, we adopt the Schrödinger Bridge framework [De Bortoli et al., 2021, Chen et al., 2021, Liu et al., 2023a, Albergo et al., 2023] as it is suitable for translation tasks where a part of the source and target samples are well aligned. We name our model *A^2SB: Audio-to-Audio Schrödinger Bridges*.

We first curate a dataset that is both expansive enough to cover most genres of music of interest and being permissively licensed. We collected and filtered permissively licensed music data from public datasets, leading to 2.3K hours in total. As data quality varies significantly across datasets, we adopt the common pre-training and fine-tuning approach [Ouyang et al., 2022].

To support both restoration tasks in a *single* model within the Schrödinger Bridge formulation [Liu et al., 2023a, Albergo et al., 2023], we frame both tasks as the generative spectrogram inpainting task: bandwidth extension as inpainting the high-frequency part of the spectrogram along the frequency

*Equal contribution. Corresponding authors: {zkong,kshih}@nvidia.com

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Music.

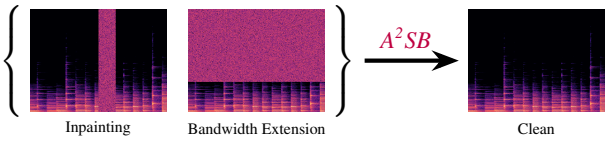


Figure 1: *A²SB* targets music restoration with a focus on inpainting and bandwidth extension, each corresponding to a specific corruption pattern in spectrogram. The model is then trained to fit the diffusion Schrödinger Bridge process from the corrupted distribution to the clean distribution.

axis, and audio inpainting as frame inpainting along the time axis. The rest of the spectrogram should exactly match the input.

Finally, we train an *end-to-end* model without using a vocoder or a codec. While prior works found success training directly on the complex spectrogram for speech enhancement [Richter et al., 2023, Jukić et al., 2024, Ku et al., 2024], we find this ineffective in our use case. We use a factorized audio representation with power compression of the magnitude and trigonometric representation of the phase. We additionally apply phase orthogonalization based on the solution of the Procrustes problem to ensure that the generated phase values are consistent. Unlike vocoder-based prior works [Liu et al., 2024, 2023b], it allows us to restore only the magnitude while preserving the original phase values.

Our model outperforms state-of-the-art baselines on several out-of-distribution test sets. We also demonstrate the effectiveness of our factorized audio representation, phase orthogonalization, and inference methods which could produce coherently restored outputs for hour-long sequences.

2 Method

Our *A²SB* is an end-to-end approach for music restoration at 44.1kHz requiring no pre-trained vocoder or audio codec. We first convert the audio to a factorized spectrogram representation (see Section 2.1). We then train a Schrödinger Bridge model for music restoration based on Liu et al. [2023a], with specific alterations for handling our audio representation (see Section 2.2). For notation, let $\tilde{X} \in [-1, 1]^L$ be the 1-D raw waveform of clean audio with length L, and X_t be the audio representation that we will use in the Schrödinger Bridge model at time t with respect to the stochastic process. We ignore the subscript t when there is no ambiguity.

2.1 Magnitude-Phase Factorized Audio Representation

The short-time Fourier transformation (STFT) representation of $\tilde{X}$, $S = \mathrm{STFT}(\tilde{X})$, is a complex matrix in $\mathbb{C}^{N \times W}$, where N is the number of frequency subbands and W is the number of overlapping STFT frames. [2] For simplicity, we can represent the complex values with their real and imaginary parts $[\mathrm{Re}(S), \mathrm{Im}(S)]$, leading to the two-channel spectrogram $S \in \mathbb{R}^{N \times W \times 2}$. While existing vocoder-free methods directly model this two-channel representation [Richter et al., 2023, Jukić et al., 2024, Wu et al., 2024], we factorize S into magnitude and phase components of the STFT in our method. We find that separating them is necessary for the following reasons: (1) magnitudes in adjacent frequency bands are strongly correlated, but this is less true for phase; (2) the periodicity of phase makes fitting to it a more challenging task; and (3) Phase-magnitude factorization isolates complications from fitting to the phase from affecting magnitude estimation. As such, we factorize S into magnitude $\Lambda_{i,j} = \sqrt{S_{i,j,1}^2 + S_{i,j,2}^2}$, and phase $\Theta_{i,j} = \mathrm{atan2}(S_{i,j,2}, S_{i,j,1})$ represented via $\cos(\Theta_{i,j})$ and $\sin(\Theta_{i,j})$, forming the final representation $X_{i,j} = \left(\Lambda_{i,j}^{\rho}, \cos(\Theta_{i,j}), \sin(\Theta_{i,j})\right)$ where our compression exponent ρ is empirically set to 0.25. Our experiments analyze the relative stability of this representation, which is also seen in works such as [Peer and Gerkmann, 2022].

2.2 Music Restoration with Schrödinger Bridges

We train a Schrödinger Bridge model on the three-channel representation X described above. Following Liu et al. [2023a], we let $X_0 \in \mathbb{R}^{N \times W \times 3}$ be the clean sample inputs, and X_1 be degraded

[2] We assume a 44.1kHz sampling rate, with hop size = 512, window length = 2048, and FFT bins = 2048. We train with $W = 256$, which corresponds to about 2.97 seconds of audio.

Table 1: Bandwidth extension results on CCMixter. See full results in Appendix B.

Method	Cutoff = 4kHz			Cutoff = 8kHz			Cutoff = 12kHz		
	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑
AudioSR	2.00	12.50	2.746	1.86	14.93	3.097	1.75	18.35	3.510
CQTDiff	2.01	14.67	1.970	2.06	15.88	1.860	2.10	16.34	1.850
IBAR	1.64	7.11	2.373	1.41	10.46	2.604	1.36	7.86	2.744
A²SB	1.85	18.00	2.851	1.62	23.39	3.438	1.45	29.26	4.211

Table 2: Inpainting results on CCMixter. See full results in Appendix C.

Method	Gap = 300ms			Gap = 500ms			Gap = 1000ms		
	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑
MAID	0.129	13.34	4.556	0.205	10.67	4.462	0.394	7.11	4.235
CQTDiff	1.305	11.16	4.486	1.293	9.01	4.403	1.266	5.95	4.126
IBAR	0.384	10.89	4.466	0.415	9.36	4.378	0.504	6.56	4.186
A²SB	0.086	15.21	4.630	0.134	12.31	4.547	0.259	8.48	4.352

Table 3: Human evaluation on bandwidth ext.

Method	MOS (bandwidth extension)		
	AAM	CCMixter	MTD
GT	4.36 ± 0.04	4.39 ± 0.05	4.26 ± 0.04
AudioSR	3.65 ± 0.08	3.67 ± 0.08	3.72 ± 0.10
CQTDiff	3.85 ± 0.08	3.10 ± 0.12	2.99 ± 0.10
IBAR	3.89 ± 0.07	2.96 ± 0.13	3.75 ± 0.07
A²SB	4.17 ± 0.06	4.17 ± 0.06	3.96 ± 0.06

Table 4: Human evaluation results on inpainting.

Method	MOS (inpainting)		
	AAM	CCMixter	MTD
GT	4.41 ± 0.05	4.36 ± 0.04	4.38 ± 0.05
MAID	3.27 ± 0.10	3.28 ± 0.10	3.33 ± 0.10
CQTDiff	3.59 ± 0.08	3.64 ± 0.09	3.63 ± 0.09
IBAR	3.70 ± 0.08	3.69 ± 0.08	3.96 ± 0.07
A²SB	4.00 ± 0.07	3.85 ± 0.08	4.09 ± 0.06

samples. We focus on bandwidth extension and inpainting, both of which can be formulated as the masking corruption similar to image inpainting. Let $\mathbb{M} \in \mathbb{B}^{N \times W \times 3}$ be the boolean mask for masking, where $\mathbb{B} = \{0, 1\}$. For bandwidth extension, $\mathbb{M}_{i,j,k} = 1$ for $i > N'$, where N' refers to the highest subband in the degraded audio. N' is randomly sampled from subbands representing frequencies above 4kHz. For inpainting, $\mathbb{M}_{i,j,k} = 1$ for $W_1 \leq j \leq W_2$, where W_1 and W_2 refer to the starting and ending frame of missing audio. Following Liu et al. [2023b], we randomly sample W_1 and W_2 such that the inpainting gap is uniform between 0.1 and 1.6 seconds. For a mask $\mathbb{M}$, we define X_1 as

$$X_1 = X_0 \odot (\mathbb{1} - \mathbb{M}) + \eta_{\text{fill}} \odot \mathbb{M}, \tag{1}$$

where $\odot$ refers to the element-wise product and $\eta_{\text{fill}} \sim \mathcal{N}(0, \sigma_{\text{fill}}^2 I)$ in order to define a Gaussian $p_{\text{deg}}(X_1|X_0)$ for the masked area in our audio representation. If $\mathbb{M} = \mathbb{1}$ and $\sigma_{\text{fill}} = 1$, the Schrödinger Bridge degenerates to an unconditional diffusion model where X_1 is Normally distributed.

3 Experiments

Baselines. For the bandwidth extension and inpainting tasks, we consider three baselines: conditional diffusion models, an inverse method, and an instruction-based method. The conditional diffusion baselines are *AudioSR* [Liu et al., 2024] for bandwidth extension and *MAID* [Liu et al., 2023b] for inpainting. The inverse method baseline is *CQTDiff* [Moliner et al., 2023], with modifications for 44.1kHz (see A.9). We train our own 44kHz instruction-based audio restoration baseline (*IBAR*) with our settings and data, given that Audit [Wang et al., 2023], the existing instruction-based model, supports only 16kHz. IBAR uses the instruction templates from Audit for both restoration tasks. Additional details provided in A.9.

Evaluation setup. We evaluate all models on several 44.1kHz out-of-distribution (OOD) test sets: AAM (synthetic music) [Ostermann et al., 2023], CCMixter (remixed music) [3] Liutkus et al. [2014] , and MTD (classical) [Zalkow et al., 2020] (see A.2 for details). Our bandwidth extension evaluation follows Liu et al. [2024] and evaluates cutoff frequencies: 4kHz, 8kHz, and 12kHz. We resample the ground truth audio to twice the cutoff frequency and use it as the input to all models. For the **inpainting evaluation**, we mask a fixed-length (300ms, 500ms, 1000ms) segment every 5 seconds. We then run the model with its receptive field centered on each masked region to inpaint the missing

[3] https://ccmixter.org/

content. For **objective evaluation metrics**, we report (1) Log-spectral distance (LSD) [Erell and Weintraub, 1990], a spectrogram distance metric; (2) Scale-invariant spectrogram-to-noise ratio (SiSpec) [Liu et al., 2021], a signal-to-noise ratio metric; and (3) ViSQOL [Chinen et al., 2020], an objective perceptual quality metric. Full details are provided in Appendix A.10. We additionally train and evaluate all models on the Maestro dataset [Hawthorne et al., 2019], and further evaluate with the F_1 score of MIDI transcriptions. We conduct **human evaluation** on the bandwidth extension (cutoff = 4kHz) and inpainting (gap = 1000ms) experiments due to the limitations of objective metrics. For each test dataset, we randomly select fifty segments and ask human listeners to rate the output quality based on how close they sound compared to the ground truth and report Mean Opinion Scores (MOS).

3.1 Bandwidth Extension Results

We show one of the bandwidth extension objective results in Table 1 and full objective results in Appendix B. The subjective results for cutoff = 4kHz are in Table 3. *A*^{2}*SB* achieves better SiSpec in most cases, indicating it has the best signal-to-noise ratio (SNR), or the least noise up to a scale transformation. Significantly better ViSQOL and MOS indicates our model has much better perceptual quality.

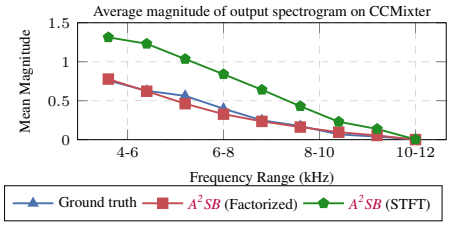


Figure 2: We compare the mean spectrogram magnitude of outputs from models trained with either the 2-channel STFT or 3-channel factorized representation, demonstrating that jointly modeling phase and magnitude without uncoupling may result in inaccurate magnitudes.

Appendix D shows qualitative samples of different bandwidth extension baselines. AudioSR often has artifacts around the cutoff frequency (Figures 9 and 10), and sometimes hallucinates implausible percussion sounds in higher frequencies (Figure 8). CQTDiff usually has worse overall quality. IBAR occasionally has incoherent generations or fails to produce a meaningful output. *A*^{2}*SB* generates better quality overall, produces coherent and consistent content, and maintains the original tempo with fewer implausible hallucinations on beats or percussions.

3.2 Inpainting Results

We show one of the inpainting objective results in Table 2 and full objective results in Appendix C. The subjective results for gap = 1000ms are in Table 4. *A*^{2}*SB* achieves consistently better evaluation results in all objective and subjective metrics. This is likely because inpainting has an easier context:synthesized-content ratio than bandwidth extension. Appendix E shows qualitative samples of different inpainting baselines, where we can see that *A*^{2}*SB* has more consistent outputs compared to that of baselines.

3.3 Necessity of Factorized Audio Representation

We find that the 3 channeled factorized representation leads to a better fit of the magnitude spectrogram than the two-channel complex representation (S in Section 2.1). In Figure 2 and Figure 19 in Appendix G, we report the average magnitude at different frequency bands. Results indicate that the complex representation poorly estimates magnitude in all frequency bands. In contrast, our three-channel factorized representation leads to similar magnitude mass compared to ground truth. We provide further qualitative analysis in Appendix G.

4 Conclusion and Future Work

This paper presents *A*^{2}*SB*, an I2SB-based novel audio restoration model for music bandwidth extension and inpainting at 44.1kHz. We present an end-to-end solution that requires no vocoder or codec, while also supporting long audio sequence processing through MultiDiffusion. We also curated a collection of permissively licensed high quality music data to train our model. Extensive experiments show that *A*^{2}*SB* achieves the state-of-the-art quality on several out-of-distribution test sets, validating the effectiveness and generalization ability of our model.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Ihab Asaad, Maxime Jacquelin, Olivier Perrotin, Laurent Girin, and Thomas Hueber. Fill in the gap! combining self-supervised representation learning with neural audio synthesis for speech inpainting. *arXiv preprint arXiv:2405.20101*, 2024.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The mtg-jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States, 2019. URL `http://hdl.handle.net/10230/42015`.

Zalán Borsos, Matt Sharifi, and Marco Tagliasacchi. Speechpainter: Text-conditioned speech inpainting. *arXiv preprint arXiv:2202.07273*, 2022.

Ricky TQ Chen and Yaron Lipman. Riemannian flow matching on general geometries. *arXiv preprint arXiv:2302.03660*, 2023.

Tianrong Chen, Guan-Horng Liu, and Evangelos A Theodorou. Likelihood training of schr\" odinger bridge using forward-backward sdes theory. *arXiv preprint arXiv:2110.11291*, 2021.

Michael Chinen, Felicia SC Lim, Jan Skoglund, Nikita Gureev, Feargus O'Gorman, and Andrew Hines. Visqol v3: An open source production ready objective speech and audio metric. In *2020 twelfth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2020.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pages 1068–1077. PMLR, 2017.

Adoram Erell and Mitch Weintraub. Estimation using log-spectral-distance criterion for noise-robust speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 853–856. IEEE, 1990.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=r1lYRjC9F7`.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ante Jukić, Roman Korostik, Jagadeesh Balam, and Boris Ginsburg. Schrödinger bridge for generative speech enhancement. *arXiv preprint arXiv:2407.16074*, 2024.

Seung-Bin Kim, Sang-Hoon Lee, Ha-Yeong Choi, and Seong-Whan Lee. Audio super-resolution with robust speech representation learning of masked autoencoder. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021.

Pin-Jui Ku, Alexander H. Liu, Roman Korostik, Sung-Feng Huang, Szu-Wei Fu, and Ante Jukić. Generative speech foundation model pretraining for high-quality speech extraction and restoration, 2024. URL https://arxiv.org/abs/2409.16117.

Junhyeok Lee and Seungu Han. Nu-wave: A diffusion probabilistic model for neural audio upsampling. *arXiv preprint arXiv:2104.02321*, 2021.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=iTtGCMDEzS_.

Sang-gil Lee, Zhifeng Kong, Arushi Goel, Sungwon Kim, Rafael Valle, and Bryan Catanzaro. Etta: Elucidating the design space of text-to-audio models. *arXiv preprint arXiv:2412.19351*, 2024.

Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin Rostamizadeh, and Ameesh Makadia. An analysis of svd for deep rotation estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22554–22565. Curran Associates, Inc., 2020.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A. Theodorou, Weili Nie, and Anima Anandkumar. I2sb: image-to-image schrödinger bridge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023a.

Haohe Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. Voicefixer: Toward general speech restoration with neural vocoder. *arXiv preprint arXiv:2109.13731*, 2021.

Haohe Liu, Woosung Choi, Xubo Liu, Qiuqiang Kong, Qiao Tian, and DeLiang Wang. Neural vocoder is all you need for speech super-resolution. *arXiv preprint arXiv:2203.14941*, 2022.

Haohe Liu, Ke Chen, Qiao Tian, Wenwu Wang, and Mark D Plumbley. Audiosr: Versatile audio super-resolution at scale. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1076–1080. IEEE, 2024.

Kaiyang Liu, Wendong Gan, and Chenchen Yuan. Maid: A conditional diffusion model for long music audio inpainting. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023b.

Antoine Liutkus, Derry Fitzgerald, Zafar Rafii, Bryan Pardo, and Laurent Daudet. Kernel additive models for source separation. *IEEE Transactions on Signal Processing*, 62(16):4298–4310, 2014.

Vincent Lostanlen and Carmine-Emanuele Cella. Deep convolutional networks on the pitch spiral for musical instrument recognition. *arXiv preprint arXiv:1605.06644*, 2016.

Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux. Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019.

Andrés Marafioti, Nathanaël Perraudin, Nicki Holighaus, and Piotr Majdak. A context encoder for audio inpainting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12): 2362–2372, 2019.

Andrés Marafioti, Piotr Majdak, Nicki Holighaus, and Nathanaël Perraudin. Gacela: A generative adversarial context encoder for long audio inpainting of music. *IEEE Journal of Selected Topics in Signal Processing*, 15(1):120–131, 2020.

Eloi Moliner and Vesa Välimäki. Behm-gan: Bandwidth extension of historical music using generative adversarial networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 943–956, 2022.

Eloi Moliner and Vesa Välimäki. Diffusion-based audio inpainting. *arXiv preprint arXiv:2305.15266*, 2023.

Eloi Moliner, Jaakko Lehtinen, and Vesa Välimäki. Solving audio inverse problems with a diffusion model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Abdulvahap Mutlu. Music instrument sounds for classification. *Kaggle*, 2024.

Vivek Sivaraman Narayanaswamy, Jayaraman J Thiagarajan, and Andreas Spanias. On the design of deep priors for unsupervised audio restoration. *arXiv preprint arXiv:2104.07161*, 2021.

Fabian Ostermann, Igor Vatolkin, and Martin Ebeling. Aam: a dataset of artificial audio multitracks for diverse music information retrieval tasks. *EURASIP Journal on Audio, Speech, and Music Processing*, 2023(1):13, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

Tal Peer and Timo Gerkmann. Phase-aware deep speech enhancement: It's all about the frame length. *JASA Express Letters*, 2(10), 2022.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel PW Ellis. mir_eval.

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

David Roberts. Piano triads wavset. *Kaggle*, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Joan Serrà, Santiago Pascual, Jordi Pons, R Oguz Araz, and Davide Scaini. Universal speech enhancement with score-based diffusion. *arXiv preprint arXiv:2206.03065*, 2022.

Chenhao Shuai, Chaohua Shi, Lu Gan, and Hongqing Liu. mdctgan: Taming transformer-based gan for speech super-resolution with modified dct spectra. *arXiv preprint arXiv:2305.11104*, 2023.

David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

Bob L Sturm. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*, 2013.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

John Thickstun, Zaid Harchaoui, and Sham M. Kakade. Learning features of music from scratch. In *International Conference on Learning Representations (ICLR)*, 2017.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Conditional flow matching: Simulation-free dynamic optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:71340–71357, 2023.

Yi-Chiao Wu, Dejan Marković, Steven Krenn, Israel D. Gebru, and Alexander Richard. Scoredec: A phase-preserving high-fidelity audio codec with a generalized score-based diffusion post-filter. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 361–365, 2024. doi: 10.1109/ICASSP48485.2024.10448371.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.

Chin-Yun Yu, Sung-Lin Yeh, György Fazekas, and Hao Tang. Conditioning and sampling in variational diffusion models for speech super-resolution. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Jun-Hak Yun, Seung-Bin Kim, and Seong-Whan Lee. Flowhigh: Towards efficient and high-quality audio super-resolution with single-step flow matching. *arXiv preprint arXiv:2501.04926*, 2025.

Frank Zalkow, Stefan Balke, Vlora Arifi-Müller, and Meinard Müller. Mtd: A multimodal dataset of musical themes for mir research. *Trans. Int. Soc. Music. Inf. Retr.*, 3(1):180–192, 2020.

A Details on Methodology

A.1 Training Datasets

In detail, the fine-tuning set includes: FMA [Defferrard et al., 2016], Medley-solos-DB [Lostanlen and Cella, 2016], MTG-Jamendo [Bogdanov et al., 2019], Musan [Snyder et al., 2015], Music Instrument [Mutlu, 2024], MusicNet [Thickstun et al., 2017], and Slakh [Manilow et al., 2019].

The pre-training set additionally includes: CLAP-Freesound [Wu et al., 2023], GTZAN [Sturm, 2013], MusicCaps [Agostinelli et al., 2023], NSynth [Engel et al., 2017], PianoTriads [Roberts, 2022].

We carefully examined all data licenses in these datasets and only selected the permissively licensed audio to train our model (i.e., we removed data that are NC, ND, SA, or under unknown licenses, etc.). As a result, the pre-training dataset has 2.3K hours and the fine-tuning dataset has 1.5K hours.

A.2 Evaluation Datasets

Here, we provide more detailed information on our evaluation datasets:

1. AAM (synthetic music) [Ostermann et al., 2023]: we randomly select 93 test samples for evaluation. Duration of samples is approximately two to three minutes.
2. CCMixter (remixed music) [4]: we use the same set as Liutkus et al. [2014]. Duration of samples is approximately between one and six minutes.
3. MTD (classical) [Zalkow et al., 2020]: we randomly select 200 test samples for evaluation. Duration of samples is between 10 seconds and 1 minute.

A.3 Architecture

Our model closely follows the conditional UNet architecture as commonly used in prior works [Ronneberger et al., 2015, Dhariwal and Nichol, 2021, Liu et al., 2023a], with some modifications. Notably, absolute positional embedding layers were replaced with 2-D rotary position embedding (RoPE) [Su et al., 2024]. Further, we use an additional conditioning variable $C \in \mathbb{R}^{N \times W}$ via absolute positional embeddings. C only varies in the frequency axis: $C_{i,j} = i, 1 \leq i \leq N$. This allows the model to strongly condition on the frequency, while maintaining translational equivariance along the temporal axis in the spectrogram.

In terms of the neural network configuration, there are five up-sampling and down-sampling layers, each having two residual blocks. The hidden channels are [128, 256, 512, 768, 1024, 2048]. Both input and output have three channels to match the 3-channel factorized representation, except in the case of the STFT baseline in 3.3. The diffusion step embedding dimension is 128, following Kong et al. [2021]. The network has 565M parameters.

A.4 Two-Stage Training

We follow the common pre-training and fine-tuning approach for stable large scale training [Ouyang et al., 2022]. During pre-training, we train our Schrödinger Bridge model from scratch on 2.3K hours of training data. We use `bf16` for more efficient training. During fine-tuning, we train on a 1.5K-hour high quality subset and use `fp32`, ensuring the model produces clean and meaningful sound for corrupted parts.

During fine-tuning, we adopt the t-range partitioning strategy from Balaji et al. [2022]: we fine-tune separate models on different t intervals, each initialized from the same pre-trained checkpoint. This leads to models specialized in different noise level ranges. We choose the intervals that partition noise level ranges between σ_0^2 and σ_1^2. In 2-partitioning, the intervals are $t \in (0, \frac{1}{2}]$ and $t \in [\frac{1}{2}, 1]$; in 4-partitioning, the intervals are $t \in (0, \frac{1}{2^{4/3}}]$, $t \in [\frac{1}{2^{4/3}}, \frac{1}{2}]$, $t \in [\frac{1}{2}, 1 - \frac{1}{2^{4/3}}]$, and $t \in [1 - \frac{1}{2^{4/3}}, 1]$. During sampling, we use the corresponding checkpoint based on the exact t.

[4] `https://ccmixter.org/`

All results presented in the main experiments section correspond to a 2-partioning, as we found diminishing returns using 4-partitioning. Full ablation testing on the impact of partitioning levels is included in the result tables later on in the appendix.

A.5 Sampling

The sampling algorithm given X_1 directly follows the diffusion model [Ho et al., 2020]. Let Δt be a step size where $\frac{1}{\Delta t}$ is an integer referring to the number of sampling steps. There is an analytic form for the posterior (see proof of Proposition 3.3 in Liu et al. [2023a]):

$$p(X_{t-\Delta t}|X_0, X_t) = \mathcal{N}\left(\frac{(\Delta\sigma_t^2)X_0 + \sigma_t^2 X_t}{\Delta\sigma_t^2 + \sigma_t^2}, \frac{(\Delta\sigma_t^2)\sigma_t^2}{\Delta\sigma_t^2 + \sigma_t^2} I\right), \tag{2}$$

where $\Delta\sigma_t^2 \coloneqq \sigma_t^2 - \sigma_{t-\Delta t}^2$. During sampling, the X_0 is replaced with the current estimate $X_0 \coloneqq X_t - \sigma_t \epsilon(X_t, t)$. Then, repeating (2) for $\frac{1}{\Delta t}$ steps yields the final output.

A.6 Long Audio Sampling with MultiDiffusion

Algorithm 1 MultiDiffusion sampling at step t

Input: $\mathbf{X}_t^{\text{full}} \in \mathbb{R}^{N \times W^{\text{full}} \times 3}$, t, W, h, $\epsilon(\cdot, \cdot)$
$\mathbf{C}, \mathbf{V} \leftarrow \mathbf{0} \in \mathbb{R}^{N \times W^{\text{full}} \times 3}$
$j \leftarrow 0$
while $j + W < W^{\text{full}}$ **do**
    $\mathbf{X}_t^{\text{patch}} \leftarrow \mathbf{X}_t^{\text{full}}[:, j : j + W, :]$
    $\mathbf{V}[:, j : j + W, :] \leftarrow \mathbf{V}[:, j : j + W, :] + \epsilon(\mathbf{X}_t^{\text{patch}}, t)$
    $\mathbf{C}[:, j : j + W, :] \leftarrow \mathbf{C}[:, j : j + W, :] + \mathbf{1}$
    $j \leftarrow j + h$
Output: $\mathbf{V} \oslash \mathbf{C}$ ▷ Element-wise divide

In practice, the audio we would like to up-sample or inpaint may be much longer than our training segment length. This is similar to the panorama generation problem in image generation, which could be solved by MultiDiffusion [Bar-Tal et al., 2023]. Inspired by their approach, we apply MultiDiffusion to extend our sampling process to variable length. Our algorithm is follows Algorithm 2 in Bar-Tal et al. [2023], where our condition is the degraded audio.

Formally, let $X_t^{\text{full}} \in \mathbb{R}^{N \times W^{\text{full}} \times 3}$ be a degraded sample of variable length that we would like to up-sample or inpaint. Our trained model $\epsilon(\cdot, t)$ can process inputs of size $N \times W \times 3$ where W corresponds to 256 STFT frames (2.97 seconds). At diffusion time t, we compute the model's output on the full sample $\epsilon(X_t^{\text{full}}, t)$ as follows. We process our input X_t^{full} with a sliding window of width W and shifting the position by a hop size $h < W$ (typically 128 for 50% overlap) until all of X_t is processed. Outputs in overlapping areas are uniformly averaged, though other weighting functions are topic of future work [Polyak et al., 2024]. Cyclic padding is used to ensure the last input window has a full temporal width of W.

We also study the GPU memory usage with MultiDiffusion enabled in our model, which we found we could make relatively efficient. We consider the bandwidth extension experiment with a cutoff frequency of 4kHz, and use the no-partitioning model to record GPU memory usage. [5] We demonstrate the results versus input audio length in Figure 3. The slope shows the memory usage from the cached vector fields in MultiDiffusion, which could be further optimized by moving them to the CPU after computing the vector field for each patch $\mathbf{int}_k$. The results indicate that our model can up-sample several minutes of audio on a common gaming GPU with ~ 10G memory and over an hour on a professional GPU with > 50G memory. We may obtain more memory reduction as well as acceleration by using TensorRT [6] and custom CUDA kernels [7].

[5] Note that for partitioned models, we could move unused checkpoints to CPU for each t-range.

[6] `https://github.com/NVIDIA/TensorRT`

[7] `https://pytorch.org/tutorials/advanced/custom_ops_landing_page.html#custom-ops-landing-page`

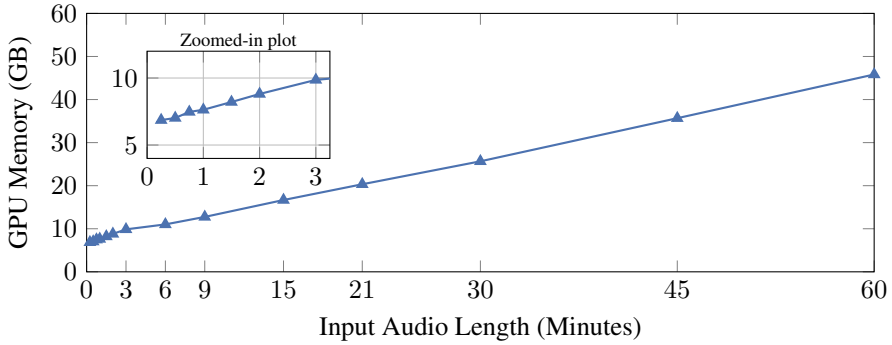


Figure 3: GPU memory usage verses input audio length (in minutes) at inference time with MultiDiffusion enabled. The GPU memory is recorded for the bandwidth extension experiment with a cutoff frequency at 4kHz. The results show that our model can up-sample several minutes of audio on a common gaming GPU and over an hour on a professional GPU.

A.7 Waveform Synthesis with Phase Orthogonalization

All operations defined in Section 2.1 are invertible and should allow us to recover the original 1-D waveform signal almost exactly. We can reconstruct the two-channel spectrogram $\hat{S}$ from X with:

$$\begin{cases} \hat{S}_{i,j,1} &= X_{i,j,2} \cdot (X_{i,j,1})^{1/\rho} \\ \hat{S}_{i,j,2} &= X_{i,j,3} \cdot (X_{i,j,1})^{1/\rho} \end{cases} . \tag{3}$$

Then, applying the inverse STFT with the same STFT parameters yields the waveform.

However, when sampling from our trained neural network, we cannot guarantee that the unconstrained model outputs $[X_{i,j,2}, X_{i,j,3}]$ satisfy the trigonometric representation of phase: $X_{i,j,2}^2 + X_{i,j,3}^2 = 1$. This could manifest as an additional scaling of the reconstructed spectrogram S, which is undesirable. To alleviate this issue, we use phase orthogonalization to map $[X_{i,j,2}, X_{i,j,3}]$ to the last-squares-nearest valid configuration. This is in part inspired by the analysis in Levinson et al. [2020] for learning 3D rotations, though we require only the 2D rotations in our case. Furthermore, the least-squares optimality of SVD orthogonalization is ideal for the removal of small amounts of Gaussian noise, making it compatible with the Gaussian diffusion process. Approaches such as Chen and Lipman [2023] can also guarantee proper roration values, but we find our approach to be simple and practical enough for our use case.

Let $\hat{R}_{i,j} \in \mathbb{R}^{2\times 2}$ be a noisy estimate of a rotation matrix at spectrogram coordinate (i, j), which is constructed with

$$\hat{R}_{i,j} \coloneqq \begin{bmatrix} X_{i,j,2} & -X_{i,j,3} \\ X_{i,j,3} & X_{i,j,2} \end{bmatrix} . \tag{4}$$

We then compute its nearest valid configuration in least squares as follows:

$$\texttt{SVDO}^+(\hat{R}_{i,j}) \coloneqq \operatorname*{argmin}_{R_{i,j}\in \mathrm{SO}(2)} \|R_{i,j} - \hat{R}_{i,j}\|_F^2, \tag{5}$$

where $\mathrm{SO}(2)$ is the orthogonal group in two dimensions. Note that for any 2×2 matrix A, we have the following solution [Levinson et al., 2020, Schönemann, 1966]:

$$\texttt{SVDO}^+(A) = U\Sigma' V, \text{ where } \Sigma' = \mathrm{diag}(1, \det(UV^\top)), \tag{6}$$

where $A = U\Sigma V$ is the SVD decomposition. Applying SVD to $\hat{R}_{i,j}$ yields

$$U = (X_{i,j,2}^2 + X_{i,j,3}^2)^{-\frac{1}{2}} \hat{R}_{i,j}, \Sigma = (X_{i,j,2}^2 + X_{i,j,3}^2)^{\frac{1}{2}} I, V = I. \tag{7}$$

And therefore, the solution is

$$\texttt{SVDO}^+(\hat{R}_{i,j}) = (X_{i,j,2}^2 + X_{i,j,3}^2)^{-\frac{1}{2}} \hat{R}_{i,j} \tag{8}$$

as $\det(UV^\top) = 1$. Then, the orthogonalized phase estimation allows us to reconstruct the spectrogram with

$$\hat{S}_{i,j} = (X_{i,j,1})^{1/\rho} \cdot (\texttt{SVDO}^+(\hat{R}_{i,j}))_{:,1}. \tag{9}$$

We further compute the minimum residual as

$$\begin{aligned}
\mathbf{Err}_{\text{phase-ortho}}(X_{i,j}) &= \|\texttt{SVDO}^+(\hat{R}_{i,j}) - \hat{R}_{i,j}\|_F^2 \\
&= \left((1 - (X_{i,j,2}^2 + X_{i,j,3}^2)^{-\frac{1}{2}}\right)^2 \|\hat{R}_{i,j}\|_F^2 \\
&= \left((1 - (X_{i,j,2}^2 + X_{i,j,3}^2)^{-\frac{1}{2}}\right)^2 \cdot 2\left(X_{i,j,2}^2 + X_{i,j,3}^2\right)^2 \\
&= 2\left((X_{i,j,2}^2 + X_{i,j,3}^2)^{\frac{1}{2}} - 1\right)^2.
\end{aligned} \tag{10}$$

We additionally note that manifold generative models such as Chen and Lipman [2023] could also address this issue without orthogonalization, yet we find our approach simple and effective enough and therefore leave this approach for future work.

A.8 Necessity of Phase Orthogonalization

We study the impact of applying phase orthogonalization, where we find that the model's output are sufficiently close to being proper rotations and require only small adjustments. In Figure 4, we visualize the distribution of the phase orthogonalization error $\mathbf{Err}_{\text{phase-ortho}}(X_{i,j})$ in (10) at different frequency bands. In detail, we consider the bandwidth extension task with cutoff = 4kHz. We take the generated part (above 4kHz) of the output spectrogram and uniformly split it into 9 bins along the frequency axis. We then plot the distribution of $\mathbf{Err}_{\text{phase-ortho}}$ values within each bin.

We note that orthogonalization error is very small (the average error is around the order of 10^{-5}), indicating that our model is able to learn the proposed audio representation very well. Only a small fraction ($< .1\%$) of the spectrogram may have larger phase orthogonalization error (up to 1.5), which will be corrected by phase orthogonalization. Overall, the phase orthogonalization provides the necessary guarantee to ensure proper STFT inversion, while likely having nominal impact on perceptual quality given the scale of its adjustments.

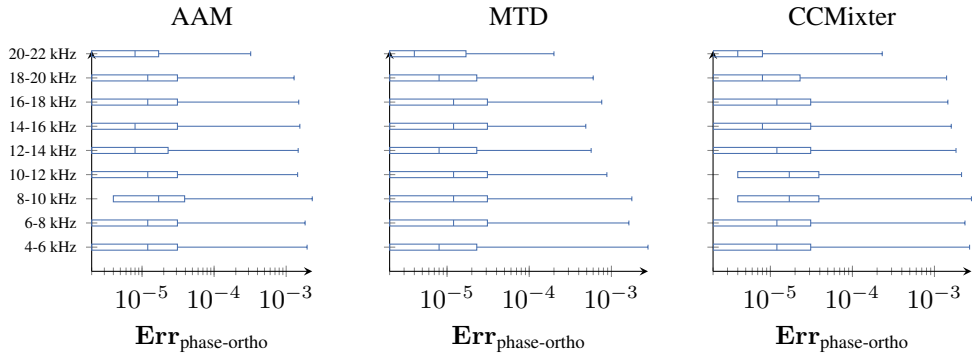


Figure 4: These box-plots visualize the distribution of the (log-scale) phase orthogonalization error $\mathbf{Err}_{\text{phase-ortho}}$ in (10) without any orthogonalization correction. The left-most whisker is omitted and is effectively zero. The right most whisker represents the 99.9-th percentile, where the outlying $.01\%$ is omitted from the graph and may have an error of as high as 1.5. The results indicate that our model can predict very accurate trigonometric values of phase for most of the time, and the phase orthogonalization acts primarily as an occasionally necessary safeguard.

A.9 Baselines

For each of the bandwidth extension and inpainting task, we consider three baselines: conditional diffusion models, inverse method, and instruction-based method. The conditional diffusion baselines are *AudioSR* [Liu et al., 2024] for bandwidth extension and *MAID* [Liu et al., 2023b] for inpainting.

The inverse method baseline is *CQTDiff* [Moliner et al., 2023]. Since it is only 22.05kHz, we re-train a larger 44.1kHz CQTDiff. We increase the depth from six to eight and double the channels to $[64, 128, 128, 256, 256, 256, 256, 256, 256]$, leading to a $5.75\times$ larger model. It is the largest model we find to have stable training in our experiments. We train our own 44kHz instruction-based audio restoration baseline (*IBAR*) with our settings and data, given that Audit [Wang et al., 2023], the existing instruction-based model, supports only 16kHz. IBAR uses the instruction templates from Audit for both restoration tasks. We use a numerically optimized diffusion transformer with adaptive layer norm [Peebles and Xie, 2023, Lee et al., 2024] that cross-attends to mel-spectrograms, the byT5 embedding [Xue et al., 2022], the OT-CFM loss function [Lipman et al., 2022, Tong et al., 2023], and a 44kHz BigVGAN-v2 vocoder [Lee et al., 2023].

A.10 Objective Evaluation Metrics

We report the following objective evaluation metrics.

- Log-spectral distance (LSD) [Erell and Weintraub, 1990], a spectrogram distance metric computed as

$$\text{LSD} = \frac{1}{W}\sum_{j=1}^{W}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\log_{10}\frac{\Lambda_{i,j}^2}{\hat{\Lambda}_{i,j}^2}\right)^2\right]^{\frac{1}{2}}, \tag{11}$$

  where Λ is the ground truth magnitude and $\hat{\Lambda}$ the magnitude of the model's prediction.

- Scale-invariant spectrogram-to-noise ratio (SiSpec) [Liu et al., 2021], a signal-to-noise ratio metric computed as

$$\text{SiSpec} = 10 \cdot \log_{10}\frac{\|\mathbf{n}(\Lambda)\|^2}{\|\mathbf{n}(\Lambda) - \hat{\Lambda}\|^2}, \tag{12}$$

  where $\mathbf{n}(\Lambda) = \langle\hat{\Lambda}, \Lambda\rangle\Lambda/\|\hat{\Lambda}\|^2$ is the scale invariant normalization of the ground truth magnitude.

- ViSQOL [Chinen et al., 2020], an objective perceptual quality metric for 48kHz audio, which measures similarity scores by comparing the spectro-temporal features and maps to the Mean Opinion Score (MOS) scale between 1 and 5. The ground truth has a score of 4.732.

- For the Maestro dataset, we further report the F_1 score of MIDI transcriptions using the `mir_eval` package [Raffel et al.].

B Full Objective Results on Bandwidth Extension

Table 5: Bandwidth extension results on AAM (synthetic music).

Method	Cutoff = 4kHz			Cutoff = 8kHz			Cutoff = 12kHz		
	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑
AudioSR [Liu et al., 2024]	2.22	13.73	3.057	1.94	14.61	3.455	1.62	19.93	3.783
CQTDiff [Moliner et al., 2023]	2.37	19.91	1.926	2.39	22.22	1.928	2.42	22.63	1.965
IBAR	1.38	8.51	2.951	1.16	10.82	3.384	0.99	12.31	4.102
A^2*SB* (no partitioning)	1.40	19.28	3.004	1.15	27.35	3.412	0.99	31.33	3.947
A^2*SB* (2-partitioning)	1.44	23.03	3.248	1.15	28.69	3.706	0.99	31.76	4.231
A^2*SB* (4-partitioning)	1.49	22.59	3.110	1.20	28.46	3.773	1.04	31.67	4.340

Table 6: Bandwidth extension results on CCMixter (remixed music).

Method	Cutoff = 4kHz			Cutoff = 8kHz			Cutoff = 12kHz		
	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑
AudioSR [Liu et al., 2024]	2.00	12.50	2.746	1.86	14.93	3.097	1.75	18.35	3.510
CQTDiff [Moliner et al., 2023]	2.01	14.67	1.970	2.06	15.88	1.860	2.10	16.34	1.850
IBAR	1.64	7.11	2.373	1.41	10.46	2.604	1.36	7.86	2.744
A^2*SB* (no partitioning)	1.93	14.05	2.770	1.71	19.95	3.200	1.48	27.17	4.047
A^2*SB* (2-partitioning)	1.85	18.00	2.851	1.62	23.39	3.438	1.45	29.26	4.211
A^2*SB* (4-partitioning)	1.84	17.46	2.657	1.65	23.17	3.430	1.50	29.20	4.234

Table 7: Bandwidth extension results on MTD (classical music).

Method	Cutoff = 4kHz			Cutoff = 8kHz			Cutoff = 12kHz		
	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑
AudioSR [Liu et al., 2024]	1.75	21.74	3.391	1.81	27.26	3.226	1.85	28.97	3.150
CQTDiff [Moliner et al., 2023]	1.74	10.62	1.747	1.63	17.42	1.777	1.57	21.62	2.000
IBAR	1.12	12.31	2.995	0.92	12.94	3.525	0.85	13.08	3.843
A^2*SB* (no partitioning)	1.33	25.51	2.557	1.05	33.10	3.201	0.87	35.34	3.936
A^2*SB* (2-partitioning)	1.29	28.15	3.101	1.07	34.36	3.718	0.88	35.97	4.200
A^2*SB* (4-partitioning)	1.77	27.56	3.446	1.59	34.25	3.829	1.51	36.07	4.274

Table 8: Bandwidth extension results on Maestro (classical piano music with MIDI).

Method	Cutoff = 4kHz			Cutoff = 8kHz			Cutoff = 12kHz		
	LSD ↓	SiSpec ↑	F_1 ↑	LSD ↓	SiSpec ↑	F_1 ↑	LSD ↓	SiSpec ↑	F_1 ↑
CQTDiff [Moliner et al., 2023]	1.154	31.49	0.761	1.137	32.99	0.772	1.129	33.33	0.774
IBAR	0.769	12.69	0.769	0.688	12.22	0.757	0.616	13.48	0.770
A^2*SB* (4-partitioning)	0.773	34.32	0.910	0.659	41.69	0.910	0.545	42.60	0.910

C Full Objective Results on Inpainting

Table 9: Inpainting results on AAM (synthetic music).

Method	Gap = 300ms			Gap = 500ms			Gap = 1000ms		
	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑
MAID [Liu et al., 2023b]	0.139	14.37	4.570	0.208	11.42	4.504	0.378	7.74	4.305
CQTDiff [Moliner et al., 2023]	1.516	14.37	4.502	1.510	11.13	4.457	1.494	7.17	4.219
IBAR	0.512	8.67	4.231	0.420	9.66	4.383	0.525	6.88	4.204
A^2SB (no partitioning)	0.081	17.10	4.660	0.128	12.72	4.592	0.257	7.90	4.432
A^2SB (2-partitioning)	0.077	17.89	4.666	0.122	13.95	4.601	0.238	9.31	4.442
A^2SB (4-partitioning)	0.076	18.36	4.673	0.121	13.96	4.613	0.238	9.16	4.465

Table 10: Inpainting results on CCMixter (remixed music).

Method	Gap = 300ms			Gap = 500ms			Gap = 1000ms		
	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑
MAID [Liu et al., 2023b]	0.129	13.34	4.556	0.205	10.67	4.462	0.394	7.11	4.235
CQTDiff [Moliner et al., 2023]	1.305	11.16	4.486	1.293	9.01	4.403	1.266	5.95	4.126
IBAR	0.384	10.89	4.466	0.415	9.36	4.378	0.504	6.56	4.186
A^2SB (no partitioning)	0.088	13.83	4.625	0.139	10.68	4.537	0.274	6.61	4.336
A^2SB (2-partitioning)	0.086	15.21	4.630	0.134	12.31	4.547	0.259	8.48	4.352
A^2SB (4-partitioning)	0.086	14.89	4.632	0.135	11.88	4.549	0.261	7.96	4.358

Table 11: Inpainting results on MTD (classical music).

Method	Gap = 300ms			Gap = 500ms			Gap = 1000ms		
	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑	LSD ↓	SiSpec ↑	ViSQOL ↑
MAID [Liu et al., 2023b]	0.139	9.86	4.406	0.223	7.29	4.285	0.430	3.79	4.044
CQTDiff [Moliner et al., 2023]	0.846	8.84	4.411	0.855	5.82	4.252	0.877	1.26	3.963
IBAR	0.293	13.62	4.136	0.306	11.71	4.109	0.346	7.93	4.030
A^2SB (no partitioning)	0.073	17.83	4.641	0.106	13.87	4.562	0.201	7.80	4.347
A^2SB (2-partitioning)	0.071	18.28	4.650	0.103	14.74	4.572	0.187	9.94	4.376
A^2SB (4-partitioning)	0.071	18.43	4.655	0.103	14.73	4.584	0.187	9.40	4.402

Table 12: Inpainting results on Maestro (classical piano music with MIDI).

Method	Gap = 300ms			Gap = 500ms			Gap = 1000ms		
	LSD ↓	SiSpec ↑	F_1 ↑	LSD ↓	SiSpec ↑	F_1 ↑	LSD ↓	SiSpec ↑	F_1 ↑
MAID [Liu et al., 2023b]	0.700	8.40	0.673	0.831	6.16	0.666	1.156	2.90	0.655
CQTDiff [Moliner et al., 2023]	0.691	12.24	0.818	0.703	8.53	0.814	0.741	4.38	0.798
IBAR	0.344	12.73	0.803	0.381	9.50	0.795	0.413	6.28	0.786
A^2SB (4-partitioning)	0.134	17.03	0.870	0.167	13.33	0.854	0.254	8.45	0.820

D More Samples on Bandwidth Extension (Cutoff = 4kHz)

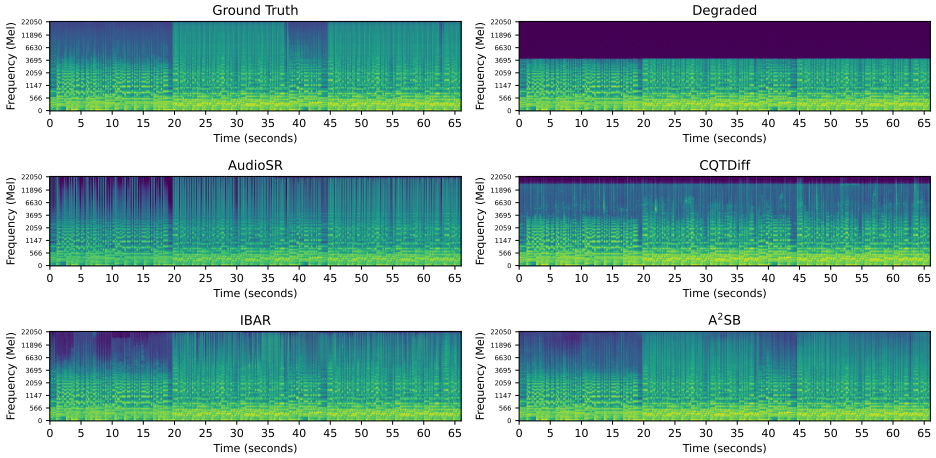


Figure 5: Qualitative comparison between different bandwidth extension methods with cutoff = 4kHz.

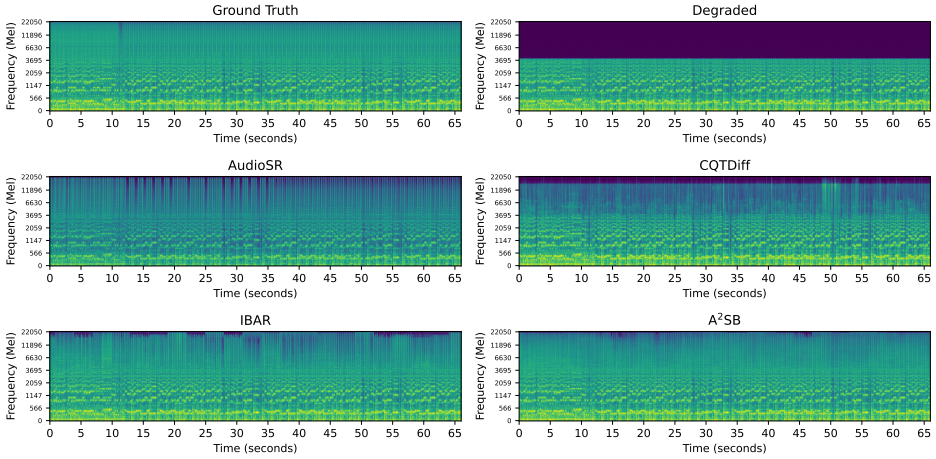


Figure 6: Qualitative comparison between different bandwidth extension methods with cutoff = 4kHz.

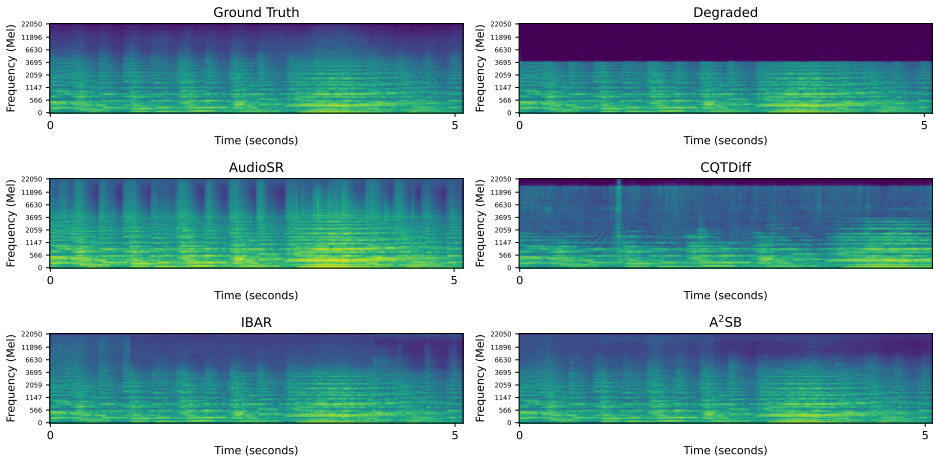


Figure 7: Qualitative comparison between different bandwidth extension methods with cutoff = 4kHz.

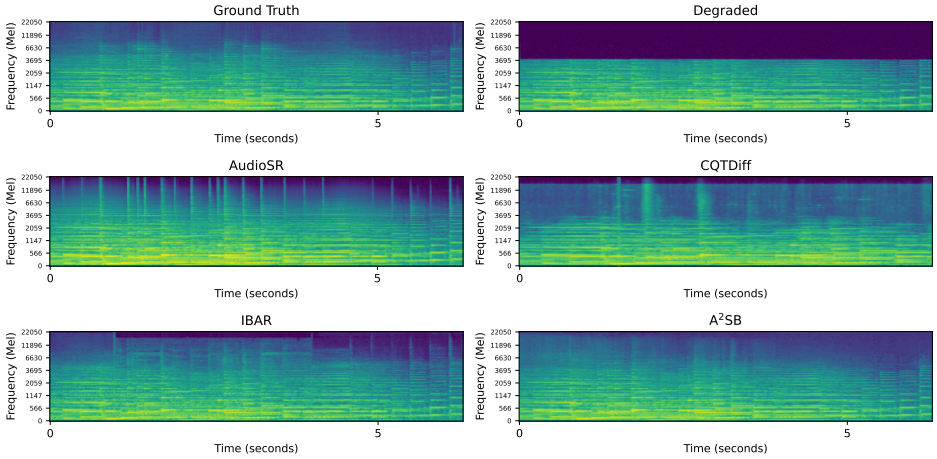


Figure 8: Qualitative comparison between different bandwidth extension methods with cutoff = 4kHz.

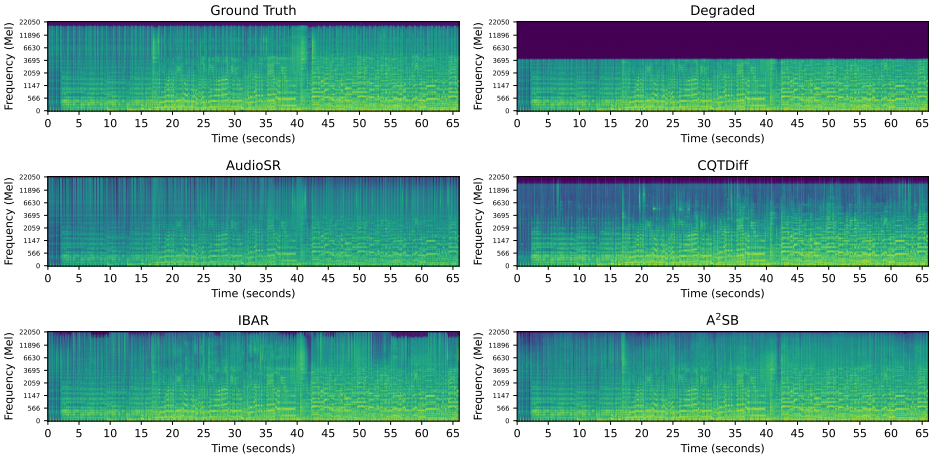


Figure 9: Qualitative comparison between different bandwidth extension methods with cutoff = 4kHz.

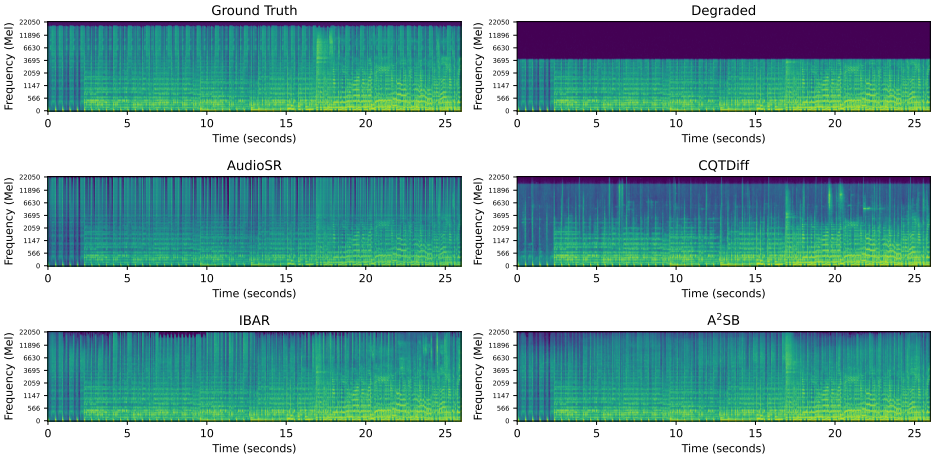


Figure 10: Qualitative comparison between different bandwidth extension methods with cutoff = 4kHz.

E More Samples on Inpainting (Gap = 1000 ms)

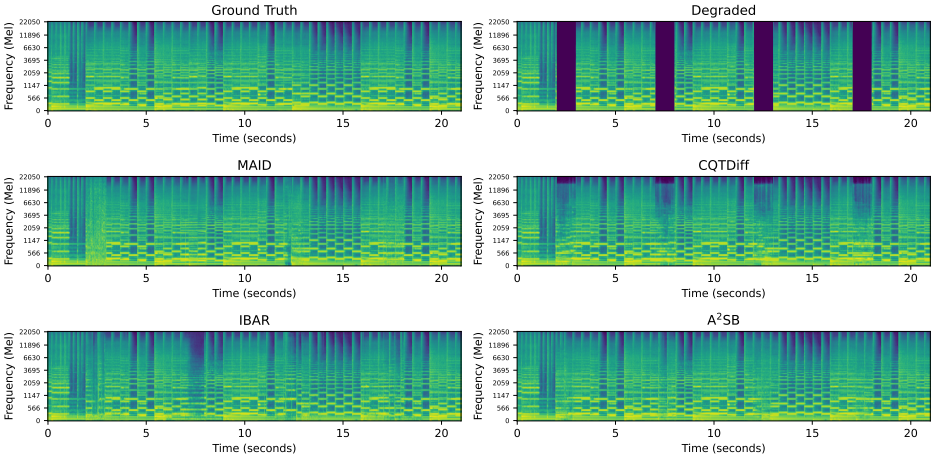


Figure 11: Qualitative comparison between different inpainting methods with inpainting gap = 1 sec.

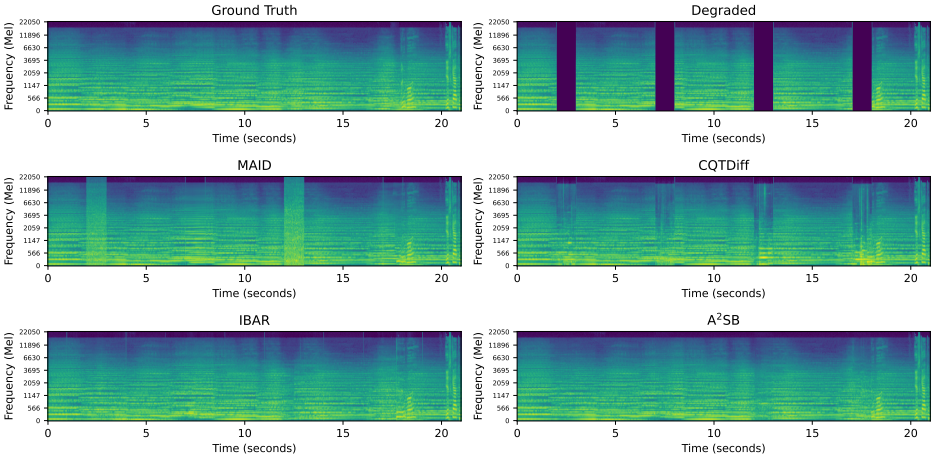


Figure 12: Qualitative comparison between different inpainting methods with inpainting gap = 1 sec.

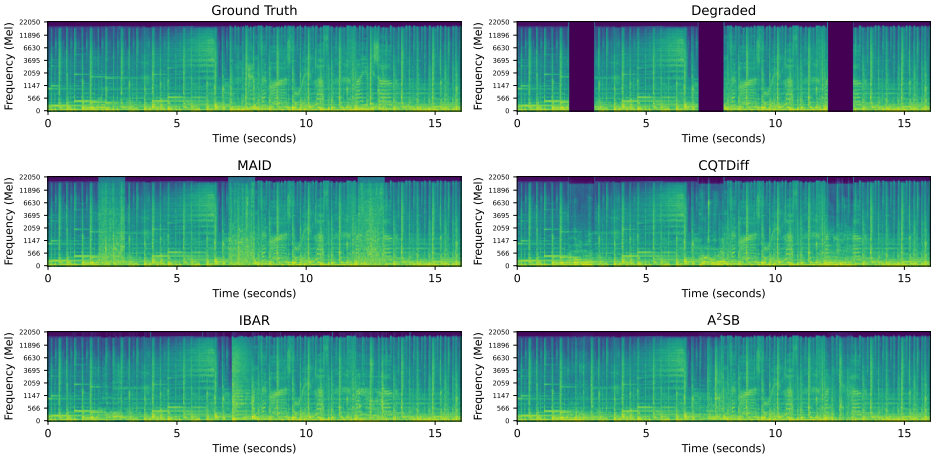


Figure 13: Qualitative comparison between different inpainting methods with inpainting gap = 1 sec.

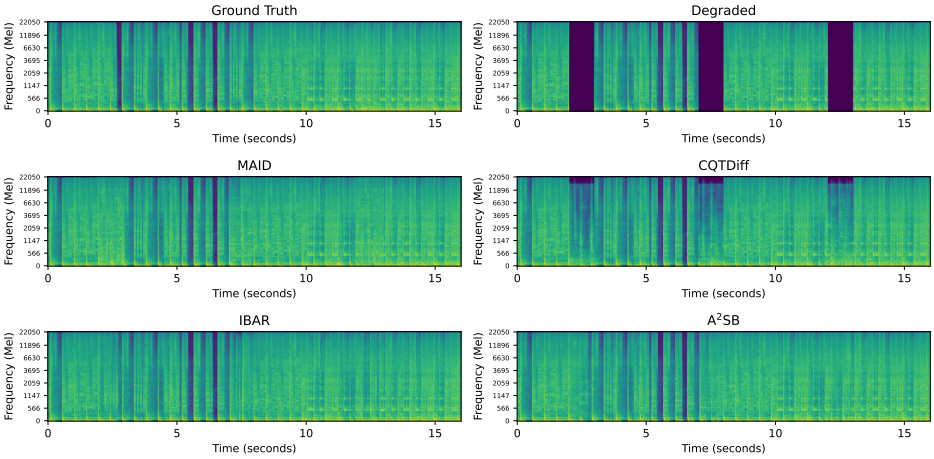


Figure 14: Qualitative comparison between different inpainting methods with inpainting gap = 1 sec.

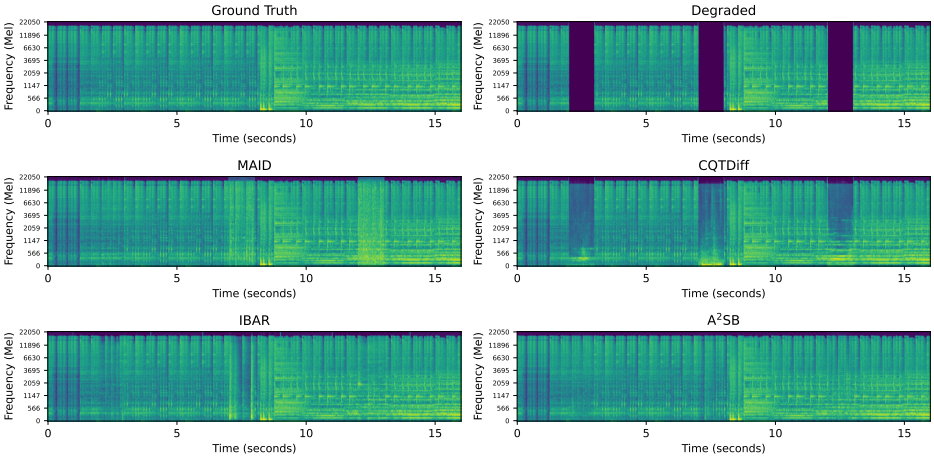


Figure 15: Qualitative comparison between different inpainting methods with inpainting gap = 1 sec.

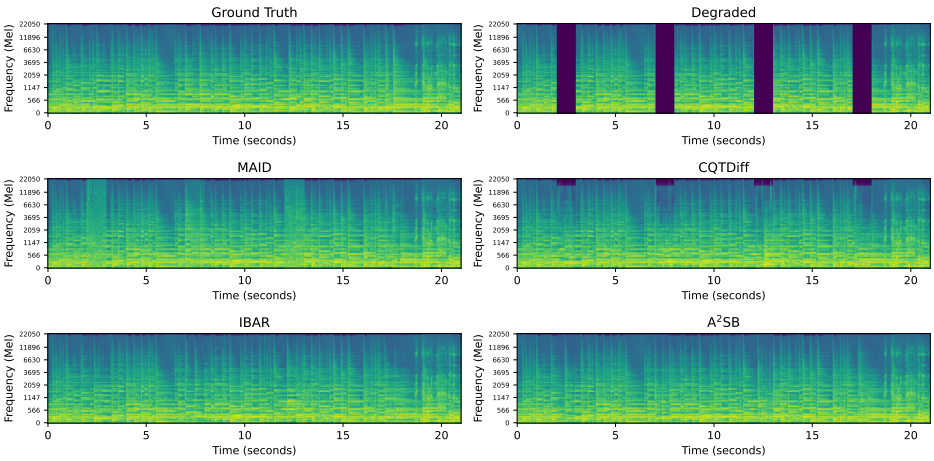


Figure 16: Qualitative comparison between different inpainting methods with inpainting gap = 1 sec.

F Additional Comparison with Audit

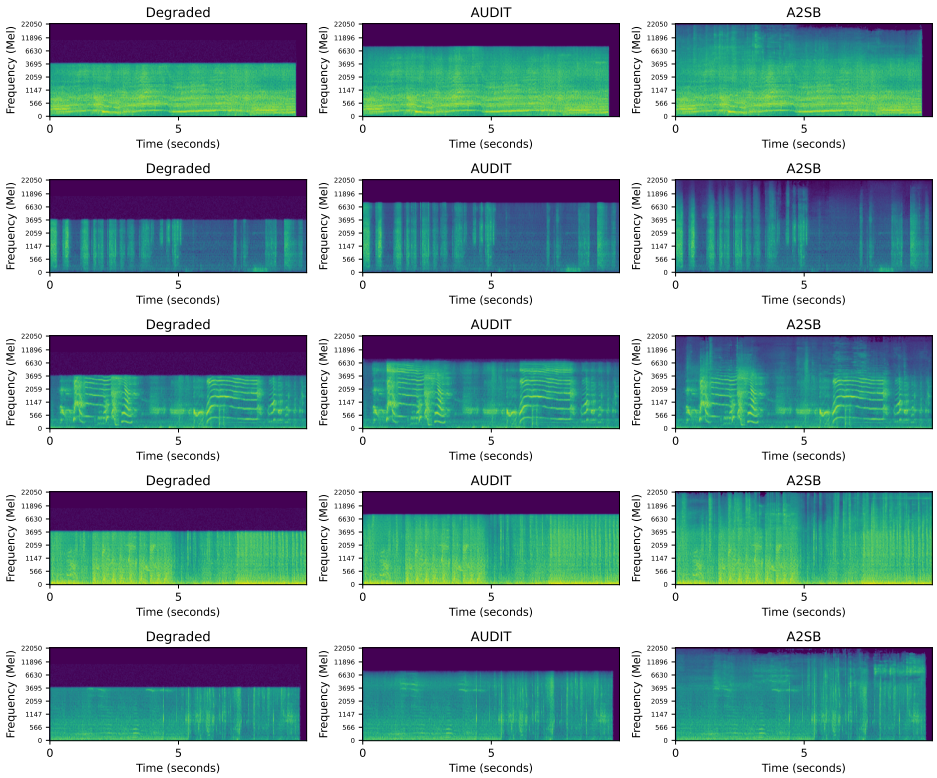
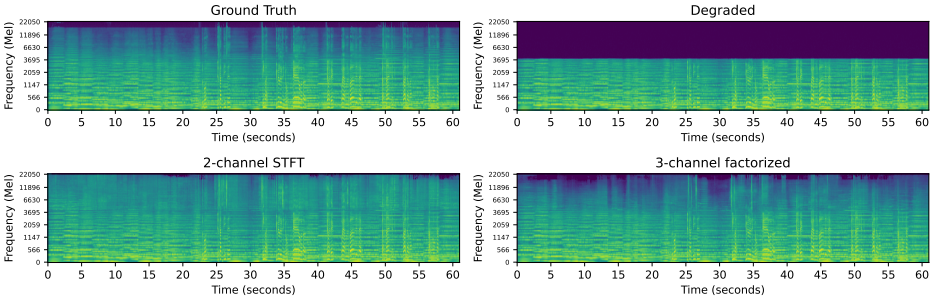
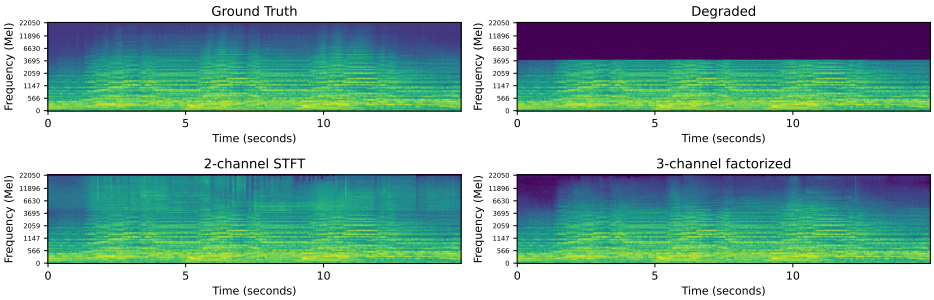


Figure 17: Qualitative comparison with Audit [Wang et al., 2023] on their demo samples for bandwidth extension. Note that these samples were not music, but rather sound effects such as car engines and gunfire. As the provided degraded samples were just under 4kHz (3.7kHz), we use the AUDIT output with a 4kHz cutoff as input to *A²SB* instead of retraining our model for a lower cutoff. Surprisingly *A²SB* demonstrates strong OOD generalization ability to non-musical sound despite having trained only on music data.

G Full Results on Necessity of Factorized Audio Representation



(a) Comparing audio representations: sample 1.



(b) Comparing audio representations: sample 2.

Figure 18: Qualitative comparison between A^2SB trained with two-channel STFT representation (S) and our proposed three-channel factorized representation. The model trained with the two-channel STFT representation has artifacts around the cutoff frequency and predicts too much content for higher frequencies, validating the effectiveness of our three-channel factorized representation.

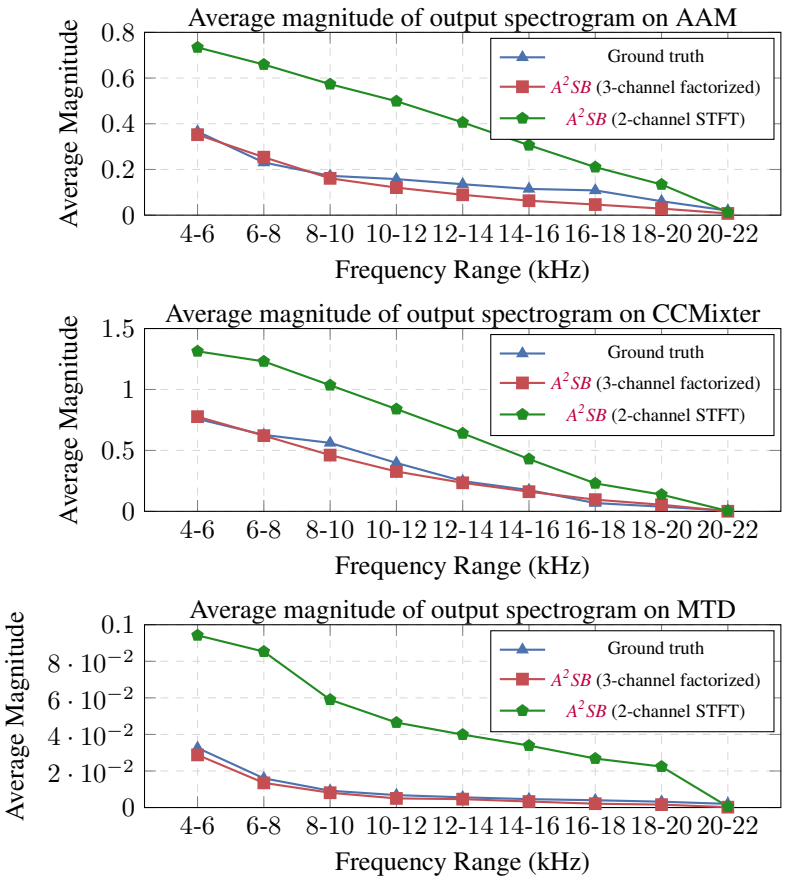


Figure 19: We compare the mean spectrogram magnitude of outputs from models trained with different audio representations: 2-channel STFT and the 3-channel factorized representation. The results demonstrate that jointly modeling phase and magnitude without uncoupling may result in overall inaccurate magnitude generations compared to that of the target distribution.

H Additional Human Evaluation Analysis

We then investigate how accurately the objective metrics could predict perceptual quality (MOS). We compute the Spearman Correlation between MOS and each objective metric ($-$LSD, SiSpec, and ViSQOL) in Table 13. Results indicate all the three objective metrics are moderately correlated with the MOS metric, but far from perfect.

Table 13: Spearman Correlation between MOS and objective metrics. All p-values are less than 0.001.

Task	**$-$LSD**	**SiSpec**	**ViSQOL**
Bandwidth extension (cutoff = 4kHz)	0.443	0.491	0.450
Inpainting (gap = 1000ms)	0.549	0.461	0.480

We additionally fit linear regression between MOS and objective metrics, and obtain the following results. For bandwidth extension,

$$\text{MOS} = 3.2158 - 0.0411 \times \text{LSD} + 0.1567 \times \text{sign}(\text{SiSpec}) \times \log|\text{SiSpec}| + 0.1015 \times \text{ViSQOL} \ (R^2 = 0.311). \tag{13}$$

For inpainting,

$$\text{MOS} = 4.1775 - 1.4022 \times \text{LSD} + 0.0681 \times \text{sign}(\text{SiSpec}) \times \log|\text{SiSpec}| + 0.0649 \times \text{ViSQOL} \ (R^2 = 0.252). \tag{14}$$

I The Effect of Sampling Steps

We study the effect of sampling steps in A^2SB. We conduct this experiment on a subset of the MTD dataset, and consider bandwidth extension with cutoff = 4kHz and inpainting with gap = 1000ms. The model is the 2-partitioning A^2SB. Results are shown in Figures 20 and 21. The results indicate that A^2SB yields almost identical generation quality with different number of sampling steps as low as 25.

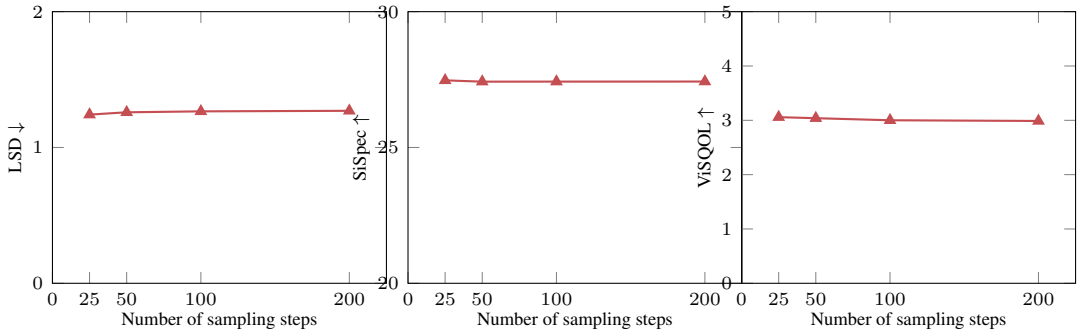


Figure 20: Objective evaluation results with different sampling steps in A^2SB. We evaluate on a subset of MTD with cutoff = 4kHz. Results indicate A^2SB has almost identical quality when we use less sampling steps.

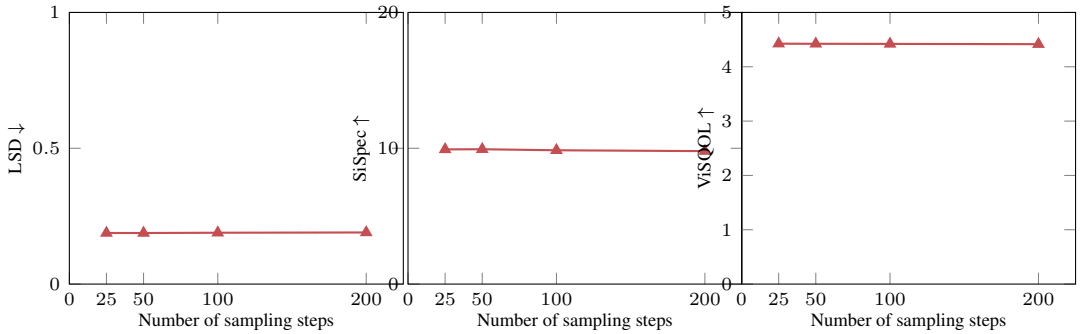


Figure 21: Objective evaluation results with different sampling steps in A^2SB. We evaluate on a subset of MTD with inpainting gap = 1000ms. Results indicate A^2SB has almost identical quality when we use less sampling steps.