Blind Audio Restoration using Contrastive Diffusion Guidance

Sattwik Basu [*1] **Chaitanya Amballa** [*1] **Zhongweiyang Xu** [1] **Jorge Vančo Sampedro** [1] **Srihari Nelakuditi** [2] **Romit Roy Choudhury** [1]

Abstract

We study blind audio restoration, where recorded signals are degraded by unknown and potentially composite corruptions. This setting is intrinsically ill-posed: many clean audio signals may be compatible with the same degraded observation. We address this ambiguity with a diffusion-based posterior sampler that generates restorations that are compatible with the observed recording. However, unlike standard generative inverse problems, the absence of a known forward model makes the likelihood score difficult to compute. Rather than estimating the degradation operator or relying on a partial approximation, we learn a contrastive embedding space in which measurement consistency can be evaluated directly. We show that the resulting embedding-space guidance provides a practical surrogate for the true likelihood score, enabling the diffusion sampler to move toward the desired posterior distribution. Experiments on historical piano recordings suggest that **AudioCoGuide** can effectively address fully blind audio inverse problems through contrastive guidance. Audio examples are available at https://audio-coguide.github.io/

1. Introduction

We consider audio restoration as a blind inverse problem: the goal is to recover a clean signal from a recording whose degradations are neither fully observed nor analytically specified. This setting is exemplified by historical music recordings, where the observed audio may contain bandwidth loss, clipping, nonlinear spectral artifacts, tape noise, room effects, and other distortions introduced by the recording medium, playback chain, or digitization process. Unlike standard restoration problems where the corruption model is known or can be reasonably parameterized, historical audio often lacks a reliable forward model. Thus, one must infer the clean signal $\mathbf{x}$ from a degraded observation $\mathbf{y}$ without knowing the operator that produced the corruption.

In the conventional inverse-problem formulation, the unknown signal $\mathbf{x} \in \mathbb{R}^m$ and the measurement $\mathbf{y} \in \mathbb{R}^l$ are related by $\mathbf{y} = \mathcal{A}(\mathbf{x}) + \mathbf{n}$, where $\mathcal{A} : \mathbb{R}^m \to \mathbb{R}^l$ denotes the forward operator and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is additive measurement noise. Given only $\mathbf{y}$, the objective is to estimate $\mathbf{x}$ by drawing samples from the Bayesian posterior $p(\mathbf{x}|\mathbf{y})$. The fundamental obstacle in using this approach is ill-posedness (Hadamard & Morse, 1953): multiple candidate signals may explain the same observation. Classical methods address this ambiguity by imposing hand-crafted priors on the structure of the signal $\mathbf{x}$, such as sparsity or total variation (Engl et al., 1996). While these priors have led to remarkable advances, they often underfit complex signal structure and require careful tuning to balance reconstruction fidelity against regularization. Recent works have shown that diffusion models (Sohl-Dickstein et al., 2015; Song et al., 2020b; Ho et al., 2020) offer a more expressive alternative by learning data-driven priors from large datasets, which can then be used for posterior sampling in inverse problems (Zheng et al., 2025; Chung et al., 2022; Song et al., 2023).

This work focuses on the fully blind case in which the degradation process is unknown at inference time. Consequently, recovering $\mathbf{x}$ requires more than a learned prior over clean audio; it also requires a mechanism for enforcing consistency with $\mathbf{y}$ without access to the true forward model.

Our approach builds on score-based posterior sampling, where the posterior score $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y})$ decomposes into a diffusion-learned prior score $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ and a likelihood score $\nabla_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x})$ (Song et al., 2020b). The prior term promotes perceptually and statistically plausible clean audio, while the likelihood term guides the sample toward agreement with the measurement $\mathbf{y}$. When $\mathcal{A}(.)$ is known and differentiable, Diffusion Posterior Sampling (DPS) (Chung et al., 2022) approximates the likelihood score using Tweedie's estimate (Efron, 2011) of the clean signal at diffusion time τ, $\hat{\mathbf{x}}_0 = \hat{\mathbf{x}}_0(\mathbf{x}_\tau) = \mathbb{E}[\mathbf{x}_0|\mathbf{x}_\tau]$, and evaluates

[*]Equal contribution [1]University of Illinois Urbana Champaign [2]University of South Carolina. Correspondence to: Sattwik Basu <sattwik2@illinois.edu>, Chaitanya Amballa <amballa2@illinois.edu>.

Accepted to the Workshop on Machine Learning for Audio, 43^{rd} International Conference on Machine Learning, 2026.

measurement consistency through an objective of the form

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau) \approx \nabla_{\mathbf{x}_\tau} \left[-\frac{1}{2\sigma_y^2} \|\mathbf{y} - \mathcal{A}(\hat{\mathbf{x}}_0)\|_2^2 \right]. \quad (1)$$

In our setting, however, this likelihood score is not directly available because the degradation operator $\mathcal{A}(.)$ is unknown.

Prior work has addressed blind audio inverse problems by estimating $\mathcal{A}(.)$ along the DPS trajectory, typically under structural assumptions on the operator, such as modeling it as a linear filter. This strategy has been effective for bandwidth extension (Moliner et al., 2024b;a), speech dereverberation (Lemercier et al., 2025; Wu et al., 2025), and source separation (Xu et al., 2025). However, extending this paradigm to broader unknown degradations is cumbersome, since it requires prior knowledge of the operator family or access to a dictionary of candidate forward models.

We take a different route. Rather than explicitly estimating the forward operator, we reformulate measurement consistency in a separate learned embedding space $\mathcal{E} \subseteq \mathbb{R}^d$. We learn this embedding space using contrastive training (Jaiswal et al., 2021; Le-Khac et al., 2020), so that clean audio and its synthetically degraded counterpart are mapped close together, while mismatched pairs are separated. The degraded examples are generated using a dictionary of forward operators, allowing the embedding to implicitly encode measurement consistency without requiring an explicit operator during inference. Importantly, this contrastive guidance is not merely a practical heuristic; it can be interpreted as a valid surrogate for the intractable likelihood score in these blind settings (Basu et al., 2025).

We instantiate our approach using the CQT-Diff framework (Moliner et al., 2024b) and its diffusion prior trained on the MAESTRO dataset (Hawthorne et al., 2018). At inference time, our method **AudioCoGuide** restores a suite of historical piano recordings affected by fully unknown degradations. Objective evaluation using Fréchet Audio Distance (FAD) (Kilgour et al., 2018) shows that **AudioCoGuide** consistently outperforms a standard signal-processing baseline based on Long-term Spectral Averaging. Although these results are preliminary, they suggest that contrastive guidance may provide a useful mechanism for fully blind diffusion-based inverse problem solving. We conclude by discussing how this formulation may generalize beyond audio restoration and outline future directions for incorporating contrastive learning into generative inverse solvers.

2. Preliminaries

Diffusion Models. Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020a;b) generate samples by reversing a gradual forward noising process. In the EDM parameterization (Karras et al., 2022), with a fixed noise schedule $\sigma(\tau) = \tau$, sampling via the probability-flow ODE (Song et al., 2020b) associated with the reverse diffusion process can be expressed as

$$d\mathbf{x}_\tau = -\dot{\sigma}(\tau)\sigma(\tau)\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)\, d\tau, \quad (2)$$

where $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$ is the score of the noisy data distribution. This score is estimated as

$$\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau) \approx \frac{s_\theta(\mathbf{x}_\tau, \tau) - \mathbf{x}_\tau}{\sigma(\tau)^2}. \quad (3)$$

where $s_\theta(\mathbf{x}_\tau, \tau)$ is a learned denoising network that provides a prior that can be used during diffusion inference.

Diffusion-Based Inverse Solvers. For inverse problems, diffusion sampling is modified to draw from the posterior rather than the unconditional prior. The posterior score decomposes via Bayes' rule as

$$\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau|\mathbf{y}) = \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau) + \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{y}|\mathbf{x}_\tau)$$

where the first term is supplied by the diffusion model and the second term enforces measurement consistency. The difficulty is that $\mathbf{y}$ is generated from the clean signal $\mathbf{x}_0$, while the sampler evolves through noisy states $\mathbf{x}_\tau$.

Diffusion Posterior Sampling (DPS) (Chung et al., 2022) handles this by approximating $p_\tau(\mathbf{y}|\mathbf{x}_\tau) \approx p(\mathbf{y}|\hat{\mathbf{x}}_0)$ with the clean estimate (i.e., $\hat{\mathbf{x}}_0 = \mathbb{E}[\mathbf{x}_0|\mathbf{x}_\tau]$) obtained from Tweedie's formula (Efron, 2011),

$$\hat{\mathbf{x}}_0 = \mathbf{x}_\tau + \sigma(\tau)^2 \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau). \quad (4)$$

When the forward operator is known and differentiable, this yields a tractable likelihood gradient for guiding reverse diffusion (Zheng et al., 2025). In our blind audio restoration problem, however, the degradation operator is unavailable, so the likelihood score cannot be computed by comparing $\mathbf{y}$ with $\mathcal{A}(\hat{\mathbf{x}}_0)$. We therefore replace operator-based likelihood guidance with a contrastive surrogate in a learned embedding space to steer the diffusion model.

3. Method

Problem Formulation. We consider blind audio restoration, where a clean waveform $\mathbf{x}$ is observed only through a degraded recording $\mathbf{y}$. The forward degradation process is modeled mathematically as

$$\mathbf{y} = \mathcal{A}(\mathbf{x}) + \mathbf{n}, \qquad \mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_y^2 \mathbf{I}), \quad (5)$$

where $\mathcal{A}$ denotes the true but unknown physical degradation operator(s). In historical audio restoration, $\mathcal{A}$ may combine bandwidth loss, clipping, spectral coloration, surface noise, compression artifacts, and other distortions. At inference time, neither the clean signal $\mathbf{x}$ nor the operator $\mathcal{A}$ is

available. The objective is therefore to recover a plausible clean waveform that is both consistent with the observed recording and likely under a learned audio prior.

Diffusion posterior samplers described in the recent literature (Chung et al., 2022; Song et al., 2023) require a likelihood score of the form $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{y}|\mathbf{x}_\tau)$ to enforce measurement consistence. When the forward operator is known, this term can be approximated by applying the operator to a denoised estimate $\hat{\mathbf{x}}_0(\mathbf{x}_\tau)$ and differentiating an audio-domain "distance" function. For example, DPS-style methods could use appropriately designed objectives of the form $C_{\text{audio}}\big(\mathbf{y}, \mathcal{A}(\hat{\mathbf{x}}_0)\big)$. Blind audio inverse solvers have extended this idea by estimating or parameterizing $\mathcal{A}$ during inference, often under structural assumptions such as linear filtering or array mixing (Moliner et al., 2024b; Xu et al., 2025; Wu et al., 2025). These approaches are effective when the degradation family is sufficiently constrained, but they become difficult to apply when the corruption is composite, nonlinear, or not easily parameterized.

We instead use the contrastive surrogate-likelihood principle to overcome this difficulty (Basu et al., 2025). The central idea is to replace explicit operator-based likelihood evaluation with a learned compatibility score between a candidate reconstruction and the observation. This is well suited to blind restoration: the missing object is precisely the likelihood term induced by an unavailable degradation operator, and a contrastively trained embedding can learn to compare clean and degraded audio without requiring $\mathcal{A}$ during the reverse diffusion process.

Contrastive Surrogate. Let $f_\varphi : \mathcal{X} \cup \mathcal{Y} \to \mathbb{R}^d$ denote an encoder with learnable parameters φ that maps both clean and degraded audio into a shared embedding space. Under the InfoNCE principle, this induces an unnormalized surrogate likelihood-to-evidence ratio

$$q_\varphi(\mathbf{x}, \mathbf{y}) \propto \exp\left(\langle f_\varphi(\mathbf{x}), f_\varphi(\mathbf{y})\rangle / \kappa\right), \tag{6}$$

where $\kappa > 0$ is a temperature parameter and $\langle \cdot, \cdot \rangle$ denotes an inner-product . The role of q_φ is not to explicitly identify the degradation operator; rather, it provides a differentiable proxy for whether a candidate clean signal could have plausibly produced the observed degraded recording.

During reverse diffusion, we form the denoised estimate $\hat{\mathbf{x}}_0(\mathbf{x}_\tau)$, and approximate the intractable likelihood score using the contrastive score in Eq. 6:

$$\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{y}|\mathbf{x}_\tau) \approx \frac{\lambda(\tau)}{\kappa} \nabla_{\mathbf{x}_\tau} \langle f_\varphi(\hat{\mathbf{x}}_0(\mathbf{x}_\tau)), f_\varphi(\mathbf{y})\rangle \tag{7}$$

Here $\lambda(\tau)$ controls the guidance strength to ensure a balance between the likelihood surrogate and the diffusion prior. The gradient is computed by backpropagating through both the denoiser s_θ and the encoder f_φ.

Because the embeddings are unit-normalized, the inner-product score is equivalent to a squared-distance objective:

$$S_\varphi(\hat{\mathbf{x}}_0, \mathbf{y}) = 1 - \frac{1}{2} \left\| f_\varphi(\hat{\mathbf{x}}_0) - f_\varphi(\mathbf{y}) \right\|_2^2 . \tag{8}$$

Thus maximizing the contrastive score is equivalent to minimizing the embedding distance between the restored candidate and the observation. Equivalently,

$$\nabla_{\mathbf{x}_\tau} S_\varphi(\hat{\mathbf{x}}_0, \mathbf{y}) = -\frac{1}{2} \nabla_{\mathbf{x}_\tau} \left\| f_\varphi(\hat{\mathbf{x}}_0) - f_\varphi(\mathbf{y}) \right\|_2^2 . \tag{9}$$

In practice, this contrastive score modifies the reverse diffusion update by replacing the unavailable operator-based likelihood term with Eq. (7). For a generic reverse ODE solver step producing an unguided proposal $\tilde{\mathbf{x}}_{\tau-h}$, we apply the guidance correction

$$\mathbf{x}_{\tau-h} = \tilde{\mathbf{x}}_{\tau-h} + \eta_\tau \nabla_{\mathbf{x}_\tau} S_\varphi\big(\hat{\mathbf{x}}_0(\mathbf{x}_\tau), \mathbf{y}\big),$$

where η_τ is the guidance step size. This update increases compatibility between the denoised estimate and the observed recording in the learned embedding space. The contrastive term can be interpreted as a tractable approximation to the likelihood score when direct likelihood evaluation is unavailable. This makes it a natural fit for blind audio restoration, where the degradation operator is not known at inference time.

Training the Contrastive Encoder. We train f_φ using paired clean and degraded piano audio from the MAESTRO dataset (Hawthorne et al., 2018). For each clean waveform $\mathbf{x}_i$, we synthesize a degraded counterpart $\mathbf{y}_i$ by applying randomly sampled corruptions in the waveform domain, including low-pass filtering, spectral coloration, clipping, additive Gaussian noise, and mild dynamic range compression. The resulting pair $(\mathbf{x}_i, \mathbf{y}_i)$ provides a positive clean–degraded example, while other samples in the mini-batch serve as corresponding negatives.

Audio is sampled at 22,050 Hz and cropped to fixed-length segments of 8.35 s. Before encoding, we compute complex Constant-Q Transforms (CQT) (Velasco et al., 2011):

$$\mathbf{X}_i = \mathrm{CQT}(\mathbf{x}_i), \qquad \mathbf{Y}_i = \mathrm{CQT}(\mathbf{y}_i).$$

We use 7 octaves and 64 bins per octave. The real and imaginary components are stacked as input channels and passed through a 2D CNN/ResNet backbone over the CQT time-frequency grid. The backbone is followed by global average pooling, a linear projection to $d = 256$ dimensions, and ℓ_2 normalization. The clean-to-degraded InfoNCE loss can be expressed as

$$\mathcal{L}_{X \to Y} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\langle f_\varphi(\mathbf{X}_i), f_\varphi(\mathbf{Y}_i)\rangle / \kappa)}{\sum_{j=1}^{B} \exp(\langle f_\varphi(\mathbf{X}_i), f_\varphi(\mathbf{Y}_j)\rangle / \kappa)}.$$

The degraded-to-clean loss $\mathcal{L}_{Y\to X}$ is defined analogously by swapping the two terms. The final symmetric contrastive objective is given by

$$\mathcal{L}_{\text{con}} = \frac{1}{2}\left(\mathcal{L}_{X\to Y} + \mathcal{L}_{Y\to X}\right). \tag{10}$$

We train the encoder with AdamW using learning rate 10^{-4}, weight decay 0, batch size $B = 256$, and temperature $\kappa = 0.07$. Training is stopped once the validation contrastive loss and pairwise alignment metrics saturate. After training, f_φ is frozen and used only to provide the contrastive guidance term in Eq. (7) during diffusion inference.

The full algorithmic pipeline for diffusion inference in **AudioCoGuide** is provided in Algorithm 1.

Algorithm 1 AudioCoGuide: Blind Audio Restoration using Contrastive Diffusion Guidance

Require: observation $\mathbf{y}$; score s_θ; contrastive encoder f_φ; guidance schedules $\{\gamma_\tau\}$; noise scales $\{\sigma_\tau\}$; step sizes $\{h_\tau\}$;

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $\tau \leftarrow T, \ldots, 1$ **do**
  $\hat{\mathbf{x}}_0 \leftarrow s_\theta(\mathbf{x}_\tau, \tau)$
  $\mathcal{L}_{\text{guide}} \leftarrow \|f_\varphi(\hat{\mathbf{x}}_0) - f_\varphi(\mathbf{y})\|_2^2$
  $\mathbf{g}_\tau \leftarrow \nabla_{\mathbf{x}_\tau} \mathcal{L}_{\text{guide}}$
  $\mathbf{s}_t \leftarrow (\hat{\mathbf{x}}_0 - \mathbf{x}_t)/\sigma_\tau^2$
  $\mathbf{x}_{\tau-1} \leftarrow \mathbf{x}_\tau - h_\tau \left[s_\tau + \gamma_\tau \cdot \mathbf{g}_\tau\right]$
return $\hat{\mathbf{x}}_0$

4. Results

Qualitative Evaluation. Fig. 1 compares spectrograms of degraded recordings, the LTAS baseline, and outputs from **AudioCoGuide** across several piano pieces. The degraded recordings exhibit missing high-frequency content, nonstationary noise, and nonlinear spectral artifacts. Compared with the original degraded audio and the LTAS baseline, our method produces restorations with visibly reduced noise and more coherent high-frequency structure. These results suggest that contrastive guidance can steer the diffusion prior toward reconstructions that are both audio-realistic and consistent with the degraded observation. Although we train on a limited dictionary of operators, empirically we found that the contrastive encoder f_φ is able to generalize across many unknown degradations in the test dataset of historical piano recordings. We provide additional audio examples on the accompanying project website to demonstrate the perceptual behavior of **AudioCoGuide**.

Quantitative Evaluation. Table 1 reports Fréchet Audio Distance (FAD) (Kilgour et al., 2018) computed using VGGNet and PANN embeddings (Kong et al., 2020). We compare the degraded recordings, the Long-term Averaged Spectrum (LTAS) baseline (Moliner et al., 2024b), and restorations produced by **AudioCoGuide**. LTAS performs spectral equalization by matching the long-term average spectrum of a degraded recording to that of a reference set. While this can correct broad spectral imbalance, it does not explicitly model nonlinear artifacts or nonstationary degradation. In contrast, **AudioCoGuide** achieves the lowest FAD under both embedding backbones, indicating that the restored samples are closer to the reference clean-audio distribution.

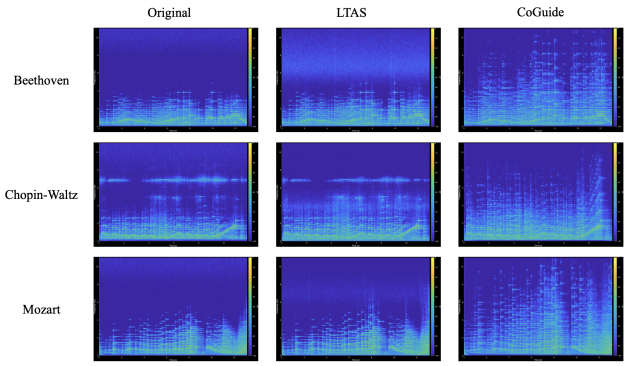

Figure 1. CQT spectrogram showing a comparison between degraded recordings, the LTAS baseline, and **AudioCoGuide** across three piano recordings. Our approach reduces broadband artifacts and recovers more high-frequency structure than the degraded input and LTAS baseline.

Audio Type	FAD-VGG ↓	FAD-PANN ↓
Degraded recording	2.52	0.39
LTAS baseline	2.88	0.27
AudioCoGuide restored	**0.84**	**0.19**

Table 1. Quantitative comparison using Fréchet Audio Distance (FAD). Lower values indicate closer agreement with the clean-audio reference distribution. FAD-PANN values are reported after scaling by 10^3.

5. Conclusions and Follow-on Work

We presented **AudioCoGuide**, a diffusion-based approach for blind audio restoration using contrastive guidance. Instead of estimating the unknown degradation operator during inference, our method uses a learned embedding-space compatibility score as a surrogate likelihood. This allows the diffusion sampler to favor restorations that remain plausible under the audio prior while staying aligned with the degraded observation. Experiments on historical piano recordings show promising qualitative improvements and lower FAD scores relative to a standard LTAS baseline. Although our experiments focus on audio restoration, the same principle can be applied to other blind inverse problems where the forward operator is unavailable, difficult to parameterize, or composed of multiple unknown degradations.

References

Basu, S., Amballa, C., Xu, Z., Sampedro, J. V., Nelakuditi, S., and Choudhury, R. R. Contrastive diffusion guidance for spatial inverse problems. *arXiv preprint arXiv:2509.26489*, 2025.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Efron, B. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Engl, H. W., Hanke, M., and Neubauer, A. *Regularization of Inverse Problems*, volume 375 of *Mathematics and Its Applications*. Springer, Dordrecht, 1996. ISBN 978-0-7923-4157-4. doi: 10.1007/978-94-009-1740-8.

Hadamard, J. and Morse, P. M. Lectures on Cauchy's Problem in Linear Partial Differential Equations. *Physics Today*, 6(8):18, January 1953. doi: 10.1063/1.3061337.

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.-Z. A., Dieleman, S., Elsen, E., Engel, J., and Eck, D. Enabling factorized piano music modeling and generation with the maestro dataset. *arXiv preprint arXiv:1810.12247*, 2018.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. A survey on contrastive self-supervised learning, 2021. URL https://arxiv.org/abs/2011.00362.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577, 2022.

Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Le-Khac, P. H., Healy, G., and Smeaton, A. F. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020. doi: 10.1109/ACCESS.2020.3031549.

Lemercier, J.-M., Moliner, E., Welker, S., Välimäki, V., and Gerkmann, T. Unsupervised blind joint dereverberation and room acoustics estimation with diffusion models. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Moliner, E., Elvander, F., and Välimäki, V. Blind audio bandwidth extension: A diffusion-based zero-shot approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:5092–5105, 2024a. doi: 10.1109/TASLP.2024.3507566.

Moliner, E., Turunen, M., Elvander, F., and Välimäki, V. A diffusion-based generative equalizer for music restoration. *arXiv preprint arXiv:2403.18636*, 2024b.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Song, J., Vahdat, A., Mardani, M., and Kautz, J. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Velasco, G. A., Holighaus, N., Dörfler, M., and Grill, T. Constructing an invertible constant-q transform with nonstationary gabor frames. *Proceedings of DAFX11, Paris*, 33:81, 2011.

Wu, Y., Xu, Z., Chen, J., Wang, Z.-Q., and Roy Choudhury, R. Unsupervised multi-channel speech dereverberation via diffusion. In *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, 2025. doi: 10.1109/WASPAA66052.2025.11231009.

Xu, Z., Fan, X., Wang, Z.-Q., Jiang, X., and Choudhury, R. R. ArrayDPS: Unsupervised blind speech separation with a diffusion prior. In *ICML*, 2025.

Zheng, H., Chu, W., Zhang, B., Wu, Z., Wang, A., Feng, B. T., Zou, C., Sun, Y., Kovachki, N., Ross, Z. E., Bouman, K. L., and Yue, Y. Inversebench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences, 2025. URL https://arxiv.org/abs/2503.11043.