SUPPORT-GUIDED OOD REMOVAL WITH DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many real-world signals are corrupted by localized, structured perturbations whose distribution is unknown: an intermittent background noise in a music recording, a watermark stamped on a photograph, a cast shadow. We study the task of recovering a clean signal x from an observation y containing such a localized out-of-distribution (OOD) corruption. We propose Support-guided OOD Removal with Diffusion, SORD, an iterative algorithm to recover x using only two ingredients: a binary map of the corruption support and a single diffusion model pretrained on clean signals. SORD does not require any paired data or any statistical model of the corruption. We evaluate SORD on natural images and human faces with spatially localized corruptions including handwritten digits, other images, watermarks, and shadows, as well as audio recordings of different musical instruments with temporally localized corruptions including other instruments and human speech. SORD is competitive with, and often outperforms, posterior-sampling inpainting given the same diffusion prior and the same mask, source-separation baselines that require a separate diffusion model for each source, and, on several benchmarks, supervised networks trained on paired clean and corrupted signals.

1 INTRODUCTION

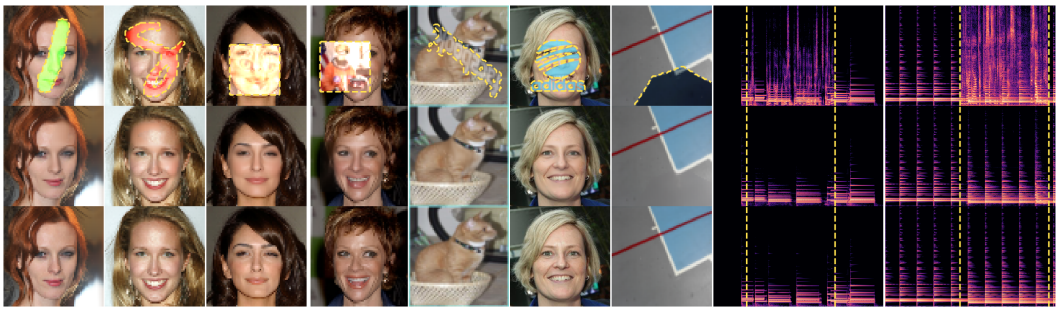

Figure 1: **SORD removes diverse localized OOD corruptions** without prior knowledge of the corruption distribution. Examples span semantic image patches, watermarks, shadows, and localized audio interference. Rows show the corrupted observation (top), our reconstruction (middle), and the clean target (bottom).

Localized corruptions arise in many restoration problems: a limited region of an image may be affected by structured noise, a watermark may cover part of an image, a cast shadow may alter a bounded region, or an interfering sound may overlap a musical recording only for a short time. These corruptions can be highly structured and semantically meaningful, rendering simple (e.g., Gaussian) noise models inadequate. At the same time, collecting paired examples or training a generative model for every possible corruption type can be impractical in the face of unexpected corruptions.

Existing approaches for removing localized corruptions generally fall into three categories. Supervised restoration methods learn a direct mapping from corrupted to clean signals (Liu et al., 2021; Liang et al., 2021; Guo et al., 2023b);these methods can perform well on specific tasks, but require paired clean and corrupted training data and are then restricted to removing corruptions similar to those seen during training. Posterior-sampling approaches combine a pretrained generative prior over clean signals with a known measurement model or likelihood (Kawar et al., 2022; Chung et al., 2023; Song et al., 2023; Wang et al., 2023). While these methods have achieved strong performance on many

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

inverse problems (Lugmayr et al., 2022; Wang et al., 2023), including inpainting a masked region, they generally require the corruption statistics to be known (Legin et al., 2023), often assuming relatively simple likelihood models such as Gaussian or Poisson noise (Chung et al., 2023) rather than corruptions with complex spatial, semantic, or temporal structure. Finally, source-separation approaches explicitly learn distributions for both the clean signal class and the corruption class, for example using separate generative models (Stevens et al., 2025; Mariani et al., 2024). These approaches can represent substantially richer corruption types, but require training data or learned generative priors for the anticipated corruption classes.

In many applications, these requirements can be restrictive or impractical. Paired clean and corrupted examples may be unavailable, and the corruption distribution may be too complex or unexpected to model accurately. We therefore consider a more weakly specified setting in which a strong generative prior is available for the clean signal class, and the corruption is localized to a known support Ω; we do not assume access to any paired corruption data, parametric corruption distribution, or generative model of the corruption class.

Our key idea is to replace statistical prior knowledge of the corruption distribution with structural prior knowledge of where or when the corruption occurs. In many applications, the corruption is localized: it affects only a spatial subset of an image, as with watermarks, shadows, or localized artifacts, or only a temporal segment of an audio signal, as when an interfering sound briefly overlaps with a recording. The corruption within this region may have complex and unknown structure, but identifying its support is often substantially easier than modeling its distribution. A broad class of anomaly-detection and segmentation methods have already been proposed to estimate the support of such localized corruptions (Defard et al., 2021; Roth et al., 2022; Wyatt et al., 2022; Zhang et al., 2023; Kheirandish et al., 2026). Figure 1 illustrates the range of Experiments in this work.

Rather than attempting to model the corruption distribution, we treat localized corruptions as out-of-distribution (OOD) relative to a generative prior over clean signals. The observations outside Ω are regarded as clean and are preserved, while the region inside Ω is reconstructed using the clean-signal generative prior. Unlike standard inpainting, the observed values within Ω are not entirely missing; instead, they contain an unknown structured corruption that may still carry information about the underlying clean signal. Our algorithm, SORD, therefore alternates between explaining the observed corrupted region and reconstructing a signal that is consistent with both the learned clean-data distribution and the trusted observations outside Ω. At each iteration, SORD uses its estimated uncertainty to adapt its effective step size and evaluate an adaptive stopping criterion. Empirically, we find that knowledge of the clean-signal prior and corruption support alone can be sufficient to match or outperform approaches that rely on paired training data or explicit corruption models.

We evaluate SORD across localized structured corruptions in both images and audio. For images, we consider both controlled additive corruptions to face images as well as specific applications in watermark removal and shadow removal in which the corruptions are not purely additive. For audio, we consider temporally localized additive audio interference in which a musical recording of one instrument is corrupted by another instrument or by human speech. Across these settings, we compare SORD with posterior-sampling (inpainting) methods using the same clean-signal prior and corruption support, supervised restoration models trained for specific corruption types, and multi-source diffusion models that explicitly model possible interfering sources. The results show that support information can substantially reduce the need for an explicit corruption model, while corruption-specific supervision remains advantageous when the test corruption is well represented by its training distribution. To summarize, our main contributions are:

1. We introduce SORD, an iterative algorithm for removing localized, highly structured OOD corruptions without requiring prior knowledge, training examples, or a generative model of the corruption. Instead, our method requires only a mask specifying the corruption support and a diffusion model trained on the clean signal distribution.

2. We demonstrate the generality of our approach across both image and audio restoration tasks, where corruptions are localized spatially or temporally. Across a diverse range of structured corruption types, SORD is competitive with approaches with substantially stronger modeling or training requirements, including supervised methods that require paired clean and corrupted data and source-separation methods that require an additional generative model trained on the corruption class.

2 Related Work

Our work lies at the intersection of diffusion-based inverse problems, corruption-specific restoration, source separation, and localized anomaly detection.

Diffusion priors for inverse problems. Diffusion models have been widely used as learned priors for inverse problems (Song et al., 2023; Chung et al., 2023; Zhu et al., 2023; Zhang et al., 2025), including inpainting a masked region (Lugmayr et al., 2022; Kawar et al., 2022; Wang et al., 2023). These methods typically assume a specified observation or measurement model together with a known or modeled noise distribution such as Gaussian noise, Poisson noise, and other non-Gaussian likelihoods (Chung et al., 2023; Legin et al., 2023). In contrast, localized structured corruptions need not follow a known parametric noise model or likelihood, motivating restoration under support information alone.

Corruption-specific image restoration. A complementary line of work trains supervised or generative restoration models for a particular family of corruptions. For example, specialized methods have been proposed for image watermark removal (Cun & Pun, 2021; Liu et al., 2021; Liang et al., 2021) and shadow removal (Guo et al., 2023b;a), using either classical supervised learning (Cun & Pun, 2021; Liu et al., 2021; Liang et al., 2021), or conditional diffusion (Guo et al., 2023b) or transformer-based (Guo et al., 2023a). However, these approaches all require paired clean and corrupted training images and are specialized to specific types of corruption, whereas we seek a single general-purpose algorithm that can remove any structured localized corruption without retraining for each corruption class.

Generative priors for source separation. A broad family of methods focuses on source separation, treating a single observation as a mixture of multiple constituent sources and using learned generative priors to recover the individual components. For images, Stevens et al. (2025) model an observation as the combination of a clean image and structured noise, using separate diffusion priors for the signal and noise distributions and jointly inferring the two components. For audio, Multi-Source Diffusion Models (MSDM) (Mariani et al., 2024) learns the joint distribution of a fixed set of musical sources, such as different instruments, and uses the resulting model for both generation and source separation. Related diffusion-based approaches learn source distributions for speech (Scheibler et al., 2023) or separately trained priors for constituent communication signals (Jayashankar et al., 2023). Despite differences in modality and model construction, these approaches rely on training data that represent the possible constituent source and corruption classes. In contrast, we assume a generative prior only for the desired clean signal; the localized corruption may come from an unknown class, may vary across observations, and need not be anticipated during training.

Localized anomaly and OOD detection. Localized anomaly and out-of-distribution (OOD) detection methods aim to identify spatial regions that deviate from the distribution of normal data. A broad range of methods estimate such localized supports using feature-space deviations (Defard et al., 2021; Roth et al., 2022; Liu et al., 2023) or learned anomaly representations (Zavrtanik et al., 2021). More recently, diffusion models have been used as expressive priors for the normal-data or inlier distribution, enabling anomalies to be localized from discrepancies produced during denoising or reconstruction (Wyatt et al., 2022; Zhang et al., 2023; He et al., 2024). These works demonstrate that the support of an anomalous or out-of-distribution component can often be estimated without explicitly modeling its distribution; our work assumes access to this corruption support and uses it towards recovering the underlying clean signal.

3 Method

We propose SORD, a diffusion-based method for removing localized structured corruptions without modeling their distribution. We assume access only to the support Ω identifying where or when the corruption occurs, and a diffusion prior pretrained on clean signals. A natural strategy under these assumptions is diffusion inpainting: discard the measurements inside Ω and generate the missing content conditioned on the uncorrupted observations outside Ω. However, such a strategy also discards potentially useful information. In our setting, the values inside Ω are not missing; rather, they contain a superposition of the clean signal and an unknown structured corruption. Our key insight is that these measurements can be progressively reused to improve an initial inpainting estimate.

SORD operationalizes this insight through three main components. First, we use diffusion inpainting as an *initialization*, and subsequently recover information from the measurements inside the corrupted

Algorithm 1 SORD: Support-Guided OOD Removal with Diffusion

Require: Observation y, corruption support Ω, corruption mask M_Ω, clean-signal diffusion denoiser D_θ, Inpainting operator R_θ, number of diffusion samples K, maximum iterations M, $\{\sigma_m\}_{m=1}^{M}$, stopping threshold τ_{stop}

Notation: $D_\theta(z, \sigma)$ denotes generic diffusion denoising of z from noise level σ, $R_\theta(y, \Omega)$ denotes inpainting operator conditioned on the clean measurements y_{Ω^c}. Our experiments instantiate D_θ via DDIM (Song et al., 2021) and R_θ via Repaint (Lugmayr et al., 2022).

for $k = 1, \ldots, K$ **do**
  $\widehat{x}_0^{(k)} \leftarrow R_\theta(y, \Omega)$, ▷ Initialize with diffusion inpainting
end for
for $m = 1, \ldots, M$ **do**
  $\sigma_m^\star \leftarrow \sqrt{\frac{1}{|\Omega|}\sum_{i\in\Omega} \mathrm{Var}\left(\{\widehat{x}_{m-1,i}^{(k)}\}_{k=1}^{K}\right)}$ ▷ Estimate uncertainty on Ω
  if $m > 1$ **and** $\sigma_{m-1}^\star - \sigma_m^\star \le \tau_{\text{stop}}\sigma_{m-1}^\star$ **then**
    return $\frac{1}{K}\sum_{k=1}^{K} \widehat{x}_{m-1}^{(k)}$
  end if
  for $k = 1, \ldots, K$ **do**
    $n_m^{(k)} \leftarrow M_\Omega \odot \left(y - \widehat{x}_{m-1}^{(k)}\right)$ ▷ Estimate additive corruption on Ω
    $\widetilde{n}_m^{(k)} \leftarrow D_\theta\left(n_m^{(k)} + \epsilon_m^{(k)}\sigma_m^\star, \sigma_m^\star\right), \ \epsilon_m^{(k)} \sim \mathcal{N}(0, I)$ ▷ Denoise the corruption estimate
    $\widetilde{x}_m^{(k)} = \begin{cases} y_i - \widetilde{n}_{m,i}^{(k)}, & i \in \Omega, \\ y_i, & i \notin \Omega. \end{cases}$ ▷ Enforce measurement consistency
    $\widehat{x}_m^{(k)} \leftarrow D_\theta\left(\widetilde{x}_m^{(k)} + \sigma_m\epsilon_m^{(k)}, \sigma_m\right), \ \epsilon_m^{(k)} \sim \mathcal{N}(0, I)$ ▷ Denoise the signal estimate
  end for
end for
return $\frac{1}{K}\sum_{k=1}^{K} \widehat{x}_M^{(k)}$

region Ω through an alternating signal–corruption decomposition. Second, because the residual between the measurement and the current reconstruction can contain both corruption and incorrectly reconstructed clean content, we use variance among multiple diffusion samples to quantify uncertainty in the current reconstruction. This uncertainty determines the scale of the subsequent diffusion-based refinement. Third, we use the same uncertainty estimate to inform an adaptive stopping criterion. The complete procedure is summarized in Algorithm 1, with key steps explained below.

We assume that the measurement consists of the clean signal x additively corrupted by an arbitrary structured corruption n within a localized region Ω:

$$y = x + M_\Omega n, \tag{1}$$

where M_Ω is a mask that restricts the corruption to its known support Ω. Importantly, we make no formal distributional assumption on n and do not require examples from its corruption class.

A straightforward approach is to ignore the corrupted measurements y_Ω entirely and perform diffusion inpainting to generate content inside Ω while preserving the trusted observations y_{Ω^c}. We initialize SORD by drawing K such inpainting samples $\{\widehat{x}_0^{(k)}\}_{k=1}^{K}$ via RePaint (Lugmayr et al., 2022), though SORD is fairly robust to this choice of inpainting method. The resulting samples are consistent with the clean-signal diffusion prior and the uncorrupted measurements y_{Ω^c}, but they need not recover the true content inside Ω. In particular, inpainting treats the corrupted region as completely unobserved even though y_Ω may still contain substantial information about x_Ω.

Our first key idea is to iteratively reintroduce this information after inpainting. At iteration m, given the current signal estimate $\widehat{x}_{m-1}^{(k)}$, we form the estimated residual

$$n_m^{(k)} = M_\Omega \odot \left(y - \widehat{x}_{m-1}^{(k)}\right). \tag{2}$$

If the current signal estimate $\widehat{x}_{m-1}^{(k)}$ were exact, this residual $n_m^{(k)}$ would equal the true corruption n. In practice, however, $\widehat{x}_{m-1}^{(k)}$ is imperfect, and consequently $n_m^{(k)}$ can contain a mixture of the true corruption and errors in the estimated clean signal. We therefore refine the raw residual before using it to update the clean signal. Let $D_\theta(z, \sigma)$ denote a generic diffusion denoising process starting from noise level σ, in our experiments we instantiate D_θ using DDIM (Song et al., 2021) with noise level σ , though SORD is again fairly robust to the choice of diffusion denoiser. We apply this diffusion refinement operator to each estimated corruption $n_m^{(k)}$:

$$\widetilde{n}_m^{(k)} = D_\theta\left(n_m^{(k)} + \sigma_m^\star \epsilon_m^{(k)}, \sigma_m^\star\right), \qquad \epsilon_m^{(k)} \sim \mathcal{N}(0, I), \tag{3}$$

using refinement scale $\sigma_m^\star$. This scale is adaptive:

$$\sigma_m^\star = \sqrt{\frac{1}{|\Omega|}\sum_{i\in\Omega} \mathrm{Var}\left(\{\widehat{x}_{m-1,i}^{(k)}\}_{k=1}^K\right)}, \tag{4}$$

and quantifies how much uncertainty remains in the current signal estimation and corresponding noise estimation. This diffusion refinement step helps to resolve fine-scale errors in the estimated corruption, even though it uses a diffusion model trained on the signal class rather than the corruption class. An ablation study is provided in Table 6 in the appendix.

We then use the refined residual $\widetilde{n}_m^{(k)}$ in conjunction with the measurements y and corruption support Ω to update our estimate of the clean signal by enforcing measurement consistency:

$$\widetilde{x}_m^{(k)} = \begin{cases} y_i - \widetilde{n}_{m,i}^{(k)}, & i \in \Omega, \\ y_i, & i \notin \Omega. \end{cases} \tag{5}$$

This measurement-informed update ensures consistency with y but does not guarantee that the resulting signal estimate $\widetilde{x}_m^{(k)}$ lies in the clean-signal distribution captured by the diffusion prior. We therefore alternate the measurement update with a diffusion-prior correction following a fixed noise schedule σ_m:

$$\widehat{x}_m^{(k)} = D_\theta\left(\widetilde{x}_m^{(k)} + \sigma_m \epsilon_m^{(k)}, \sigma_m\right), \qquad \epsilon_m^{(k)} \sim \mathcal{N}(0, I). \tag{6}$$

SORD yields stable performance across a wide range of noise schedules σ_m; we provide a sensitivity study in Table 7 in the appendix.

SORD alternates between two complementary sources of information: the *measurement*, which reveals what remains unexplained by the current estimates of the signal and the corruption, and the *clean diffusion prior*, which constrains the reconstructed signal toward plausible clean signals and refines details in the estimated corruption. Iterating these updates allows information initially discarded by inpainting to be progressively recovered.

SORD terminates adaptively once the uncertainty estimate $\sigma^\star$ converges. When this condition is satisfied, independent clean-signal hypotheses agree sufficiently within the corrupted support, and SORD returns their mean. SORD continues its alternating updates until either this condition is satisfied or it reaches a maximum of M iterations. Across our experiments, we use $M = 6$ and find that SORD terminates in fewer steps in almost all examples.

4 Experiments

We evaluate SORD on localized structured corruption removal across both image and audio domains. For images, we consider additive localized structured corruptions, visible-watermark removal, and shadow removal; for audio, we consider temporally localized additive interference removal. We report peak signal-to-noise ratio (PSNR) over the corrupted support Ω for image experiments, since full-image metrics can be dominated by unchanged pixels outside the corrupted region, and scale-invariant signal-to-distortion ratio improvement (SI-SDRi) within the interference window for audio experiments. You can find examples of every experiment in Figure 2.

4.1 Additive Localized Structured Image Corruptions

We first consider additive localized corruptions on CelebA images (Liu et al., 2015), constructed by inserting natural-image patches (van der Walt et al., 2014), MNIST digits (LeCun et al., 1998), and CIFAR-10 patches (Krizhevsky, 2009). These experiments test the target setting of this work: the corruption is additive, localized, and highly structured, while the clean diffusion prior is trained only on CelebA. Visual and quantitative results are reported in Figure 2 and Table 1, respectively.

We compare against several DPS (Chung et al., 2023) baselines using the same CelebA clean prior. *DPS-Inpaint* treats the corrupted support as missing and inpaints it by conditioning the sampling process on the observations outside Ω. We additionally evaluate DPS under a Gaussian likelihood assumption within Ω, either using the oracle corruption mean μ or fixing $\mu = 0$, since in practice the corruption mean is inaccessible. The limited performance of these variants validates that an explicit Gaussian likelihood is insufficient to remove highly structured localized corruptions. We also compare against SGM (Stevens et al., 2025), which uses separate diffusion models for the clean image distribution, CelebA, and each corruption distribution, as well as a variant of SGM which we endow with additional knowledge of the corruption support Ω. SORD outperforms both variants despite lacking any prior knowledge of the corruption distribution. Further experimental details are provided in Section A.1.

4.2 Additive Localized Audio Interference

To evaluate whether the same support-guided inference principle extends beyond images to localized audio interference; we use the Slakh2100 (Manilow et al., 2019) dataset of instrumental music and construct additive mixtures in which an interfering source is present only within a known temporal window Ω. We compare SORD with Multi-Source Diffusion Models (MSDM) (Mariani et al., 2024). MSDM performs source separation using multiple source-specific diffusion priors, whereas SORD uses only a single diffusion prior of the target source together with the temporal support Ω of the interference. The audio mixture is constructed as

$$y = x + g \cdot (\mathbf{M}_\Omega \odot s)\,, \tag{7}$$

where x is the target source and s is the interfering source. We adjust g to vary the interference-window SIR between 0 and 5 dB, and consider the default Slakh mixing level $g = 1$.

We consider both *in-vocabulary* (IV) and *out-of-vocabulary* (OOV) interfering sources, relative to the MSDM baseline. For a given target source among the MSDM-modeled sources $\{$piano, guitar, bass, drums$\}$, the IV set consists of the other three MSDM-modeled sources, while the OOV set consists of the remaining eight Slakh instrument classes for which MSDM has no corresponding diffusion model. We additionally consider human speech interference both as another OOV source for vanilla MSDM and as an IV source for an augmented version of MSDM endowed with a speech diffusion model. Quantitative results are presented in Table 1, with per-source results in Table 9 and representative spectrogram visualizations in Figure 3 in the appendix. We find that MSDM can outperform SORD for certain target instruments and when the interferer is a modeled source, but that SORD is generally more robust to unexpected interferer types as it does not rely on prior knowledge of the corruption class. Results for all target sources are provided in Table 9 in the appendix.

4.3 Watermark Removal

We next consider visible watermarks, which provide a realistic localized structured corruption containing text or logos with varying color, shape, scale, and transparency. We use watermark examples from the Colored Large-scale Watermark Dataset (CLWD) (Liu et al., 2021).

A watermarked observation follows the alpha-compositing model

$$y = (1 - \alpha) \odot x + \alpha \odot w, \tag{8}$$

where x is the clean image, w is the watermark appearance, and α is its opacity map, which may vary at the pixel level. We define the corruption support as all pixels with nonzero watermark opacity. SORD and DPS use only a diffusion prior for the clean image. Supervised watermark-removal baselines PatchWiper (Mo et al., 2025) and SLBR (Liang et al., 2021), in contrast, are trained directly on paired watermarked and clean examples.

Importantly, although we assume knowledge of the alpha-compositing form in Equation (8), we do not provide the true pixel-wise opacity map α to SORD at inference time. This deliberately introduces model mismatch, since the true watermark opacity varies spatially across the logo.

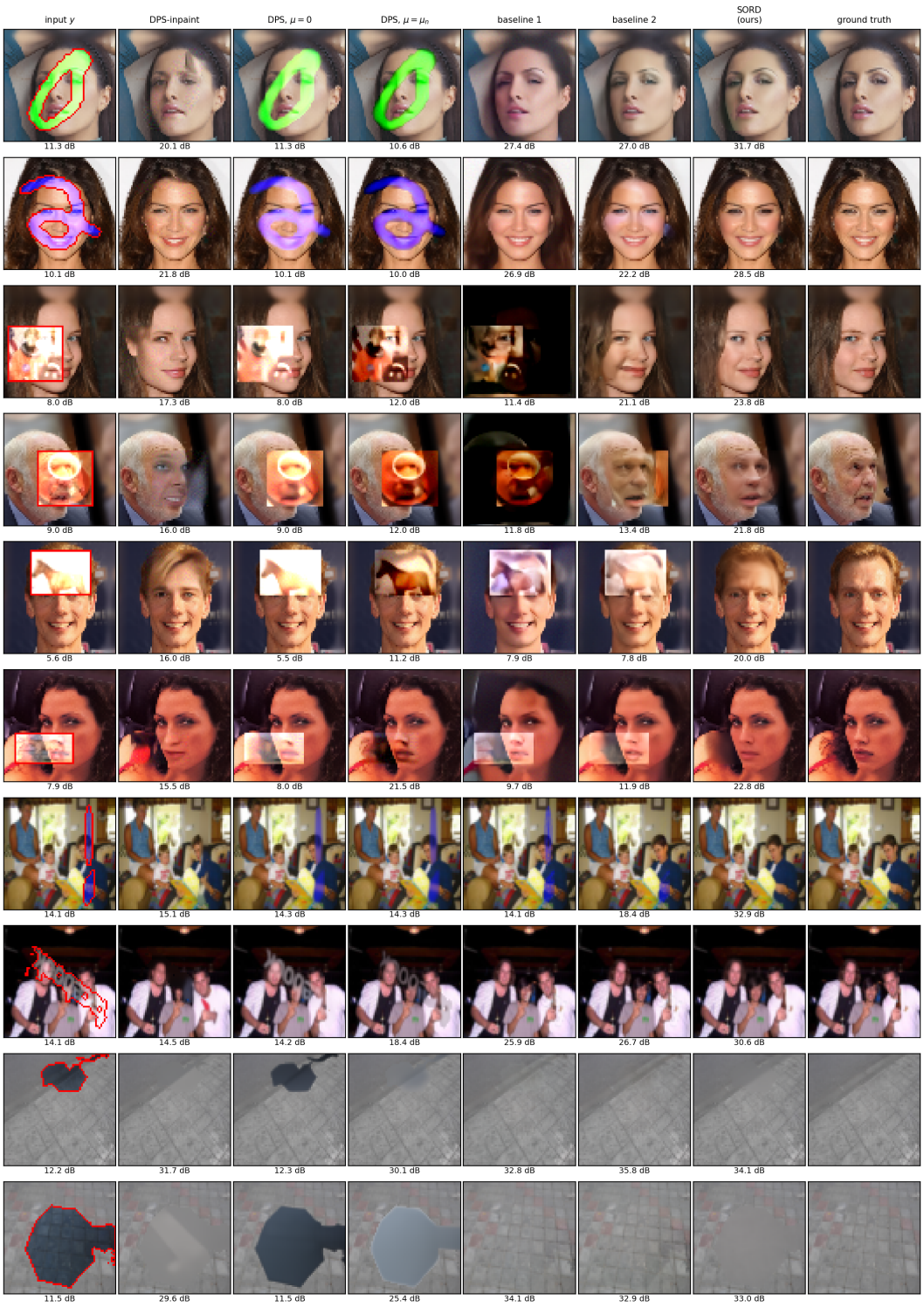

Figure 2: **Visual results on localized image corruption removal.** Each pair of rows corresponds to a different restoration setting: Rows 1–6 respectively show localized MNIST-digit, natural-image patch, and CIFAR-10 patch corruptions on CelebA faces. For those settings, Baseline 1 is Stevens et al. (2025), using separate CelebA and corruption-patch diffusion priors, while Baseline 2 is its mask-projected variant. Rows 7–8 show watermark removal on CLWD, with PatchWiper (Mo et al., 2025) as Baseline 1 and SLBR (Liang et al., 2021) as Baseline 2. Rows 9–10 show shadow removal on ISTD, with ShadowDiffusion (Guo et al., 2023b) as Baseline 1 and ShadowFormer (Guo et al., 2023a) as Baseline 2. Red contours or bounding boxes indicate the known corruption support Ω. Numbers below the reconstructions denote PSNR computed only over Ω.

Table 1: **Localized structured corruption removal across image and audio domains.** Image results report PSNR (dB) within Ω on 100 CelebA images at 64×64 resolution. Audio results report SI-SDR improvement (SI-SDRi, dB) on Ω for localized instrumental and speech interference. All values are mean $\pm$ standard deviation. IV and OOV denote in-vocabulary and out-of-vocabulary interferers for MSDM, respectively; for speech, IV uses MSDM augmented with a speech prior, while OOV uses the original four-source MSDM without a speech prior. The setting $g = 1$ denotes the unnormalized mixture. SGM and MSDM use explicit contaminant/source priors, whereas SORD uses only the clean source prior. DPS $\mu = \mu_n$ uses the oracle per-image mean of the true corruption on the support. The highest mean in each row is shown in **bold**. n denotes the number of evaluated mixtures.

Image Corruption	Input	DPS Inpaint	DPS $\mu = \mu_n$	DPS $\mu = 0$	SGM 2 priors	SGM 2 priors + mask	SORD
MNIST	12.09 $\pm$2.47	19.85 $\pm$2.77	11.06 $\pm$1.18	11.91 $\pm$2.17	25.78 $\pm$3.42	22.51 $\pm$3.89	**27.16** $\pm$2.04
Natural patch	9.37 $\pm$1.62	16.84 $\pm$2.94	15.74 $\pm$2.75	9.32 $\pm$1.58	11.45 $\pm$2.11	16.96 $\pm$3.32	**21.96** $\pm$2.32
CIFAR-10 patch	7.92 $\pm$1.90	17.28 $\pm$3.92	14.87 $\pm$2.18	7.83 $\pm$1.89	11.66 $\pm$2.83	11.25 $\pm$2.77	**22.05** $\pm$3.44

Audio Interferer	Mix	n	DPS Inpaint	DPS $\mu = \mu_n$	DPS $\mu = 0$	MSDM	SORD
IV instrument	SIR 0	60	-0.34 $\pm$6.17	-0.01 $\pm$0.05	-0.06 $\pm$0.43	**5.32** $\pm$7.64	1.12 $\pm$5.11
IV instrument	SIR 5	60	-0.12 $\pm$5.67	0.00 $\pm$0.07	0.01 $\pm$0.09	1.40 $\pm$6.85	**2.74** $\pm$6.23
IV instrument	$g = 1$	60	0.34 $\pm$0.92	-0.02 $\pm$0.13	-0.02 $\pm$0.12	**6.48** $\pm$9.06	1.29 $\pm$2.58
Speech (IV)	SIR 0	160	1.02 $\pm$1.18	-0.01 $\pm$0.10	0.00 $\pm$0.05	**8.25** $\pm$8.64	5.51 $\pm$3.37
Speech (IV)	SIR 5	160	1.85 $\pm$2.02	-0.03 $\pm$0.28	0.00 $\pm$0.05	2.00 $\pm$10.03	**6.40** $\pm$3.39
Speech (IV)	$g = 1$	160	0.94 $\pm$1.42	0.00 $\pm$0.02	0.00 $\pm$0.02	**9.74** $\pm$9.73	3.74 $\pm$3.33
OOV instrument	SIR 0	160	-0.43 $\pm$17.01	-1.14 $\pm$14.35	-1.22 $\pm$15.50	-0.24 $\pm$16.94	**1.15** $\pm$17.54
OOV instrument	SIR 5	160	0.03 $\pm$17.15	-1.21 $\pm$15.50	-1.13 $\pm$14.35	-2.67 $\pm$18.39	**1.97** $\pm$17.54
OOV instrument	$g = 1$	160	-0.34 $\pm$17.08	-1.23 $\pm$15.50	-1.23 $\pm$15.50	-1.54 $\pm$17.40	**0.79** $\pm$17.41
Speech (OOV)	SIR 0	105	1.14 $\pm$1.35	-0.01 $\pm$0.12	0.00 $\pm$0.06	3.57 $\pm$6.01	**5.71** $\pm$3.60
Speech (OOV)	SIR 5	105	2.06 $\pm$2.29	-0.04 $\pm$0.33	0.01 $\pm$0.04	-0.05 $\pm$6.80	**6.44** $\pm$3.40
Speech (OOV)	$g = 1$	160	0.94 $\pm$1.42	0.00 $\pm$0.02	0.00 $\pm$0.02	**4.15** $\pm$6.59	3.74 $\pm$3.33

We evaluate watermark removal both on the task-specific CLWD dataset and on images from FFHQ (Karras et al., 2019) and Imagenette, a ten-class subset of ImageNet (Deng et al., 2009), to which we synthetically add CLWD watermarks following Equation (8). For each dataset, PatchWiper is evaluated zero-shot (as it was pretrained on a large and diverse dataset), SLBR is trained on paired watermarked and clean images, and SORD uses a diffusion model trained only on the clean images without any prior knowledge of watermarks.

Quantitative results are reported in Table 2, with representative examples included in Figure 2. In this setting where the watermarking model (Equation (8)) deviates from the equal-weight additive corruption model assumed by SORD, SORD is competitive with, but does not always outperform, these domain-specific baselines trained on paired clean and corrupted data.

4.4 Shadow Removal

Shadows constitute a qualitatively different corruption from the additive perturbations considered above, and thus a further model-mismatch stress test of SORD. We model a shadowed image as a multiplicative corruption:

$$y = x \odot s, \tag{9}$$

where x is the clean image and s is a spatially varying attenuation field with $s < 1$ inside the shadow region Ω and $s = 1$ on Ω^c. To apply its additive corruption-removal procedure, SORD applies its measurement-consistency updates in the log-intensity domain, $\log y = \log x + \log s$, but still applies its pretrained clean-image diffusion model in the image domain (to estimates of x and s, not their pixel-wise logarithms).

We evaluate shadow removal on the Image Shadow Triplets Dataset (ISTD) (Wang et al., 2018), which provides shadowed images, corresponding shadow-free targets, and binary shadow masks. Since the shadow and shadow-free images are from separate physical camera captures, they exhibit exposure and illumination differences even outside the annotated shadow region. We further evaluate the same shadow-removal task on FFHQ and Imagenette, by synthetically applying ISTD shadows to FFHQ and Imagenette images following Equation (9). Further details are provided in Section A.1.

Table 2: **Localized structured corruption removal on watermark and shadow benchmarks.** Task-specific baselines use dedicated training data or corruption models, whereas SORD uses only the clean-image prior and the known corruption support. DPS $\mu = \mu_n$ uses the oracle per-image mean of the true corruption on the support. Highest mean in each row is shown in **bold**.

Benchmark	Input	DPS Inpaint	DPS $\mu = \mu_n$	DPS $\mu = 0$	SLBR	PatchWiper	SORD
FFHQ	13.49 ± 3.20	27.13 ± 3.81	18.80 ± 5.11	14.35 ± 3.89	18.90 ± 5.10	23.44 ± 5.59	$\mathbf{27.80 \pm 3.50}$
CLWD	22.55 ± 5.33	17.14 ± 4.74	24.26 ± 5.15	22.73 ± 5.30	$\mathbf{31.55 \pm 4.47}$	31.37 ± 6.84	28.06 ± 6.48
Imagenette	13.64 ± 4.30	24.81 ± 7.56	19.63 ± 7.18	14.56 ± 4.61	21.66 ± 5.08	21.79 ± 6.89	$\mathbf{26.43 \pm 6.56}$
Benchmark	**Input**	**DPS Inpaint**	**DPS $\mu = \mu_n$**	**DPS $\mu = 0$**	**Shadow Diffusion**	**Shadow Former**	**SORD**
ISTD	14.48 ± 3.70	27.44 ± 7.08	27.60 ± 4.21	14.57 ± 3.77	$\mathbf{30.60 \pm 4.93}$	30.49 ± 5.18	29.10 ± 5.55
FFHQ	14.89 ± 4.13	17.39 ± 3.85	19.85 ± 3.80	14.83 ± 4.09	19.15 ± 3.23	$\mathbf{23.29 \pm 2.89}$	20.34 ± 4.19
Imagenette	14.30 ± 4.71	15.97 ± 6.55	19.51 ± 4.96	14.12 ± 4.59	18.57 ± 2.83	$\mathbf{21.79 \pm 4.32}$	19.05 ± 6.50

We compare SORD against the same DPS-based restoration baselines and two task-specific supervised shadow-removal methods: ShadowDiffusion (Guo et al., 2023b), which incorporates an explicit conditional shadow degradation prior within a diffusion framework, and ShadowFormer (Guo et al., 2023a), a transformer-based method that uses global context between shadow and non-shadow regions for restoration. For each dataset, SORD uses only the corresponding clean-image prior and the shadow support, ShadowDiffusion is trained with paired clean and shadowed examples from each dataset, and ShadowFormer is applied zero-shot (it was pre-trained on a huge dataset).

As shown in Table 2, the log-domain formulation performs comparably to the supervised ShadowDiffusion baseline within the shadow support, despite the handicap of using no shadow-specific training data, while ShadowFormer has slightly stronger performance.

4.5 Ablation and Sensitivity Studies

We study the sensitivity/robustness of SORD to its main hyperparameters and design choices, with full results in the appendix. Specifically, we examine the effect of the number of samples K in Section A.2, the noise schedule σ_m in Table 7, the choice of D_θ for diffusion sampling and its cost in Section A.3, the relative value of statistical versus support-based corruption modeling in Table 8, and the robustness of SORD to imperfect corruption masks in Section A.4. Finally, because our method refines the estimated corruption by passing it through the clean-signal diffusion prior at a small noise level, Table 4 evaluates the effectiveness of this diffusion n-refinement and its effect on the final reconstruction quality.

5 Discussion

We introduced SORD for removing localized structured corruptions using only a pretrained clean-signal diffusion prior and knowledge of the corruption support, without requiring paired clean/corrupted data or a model of the corruption distribution. Across image and audio tasks, this support-guided strategy consistently improves over simply discarding the corrupted measurements, as is common in inpainting, and is competitive with methods that rely on substantially stronger assumptions, including corruption-specific supervision or additional source priors.

Our experiments also highlight the tradeoffs of this weakly specified setting. When the corruption class *is* known and well represented by training data, methods that explicitly model that class can retain an advantage; conversely, SORD is particularly useful when the corruption is unexpected or otherwise difficult to model. SORD also depends on a sufficiently accurate corruption support: overestimating the support tends to degrade performance gracefully, whereas leaving corrupted pixels outside the support can be substantially more harmful because those observations are treated as trustworthy. Future work may focus on jointly estimating the corruption support alongside corruption removal, developing more principled uncertainty-guided refinement and stopping criteria, and extending the framework to broader non-additive or partially specified corruption models.

Reproducibility Statement

We provide detailed information to support reproducibility of our results. The main paper describes the proposed method and experimental protocol, while the appendix provides additional implementation and experimental details as well as ablation and sensitivity studies. Code will be released upon publication.

AI Use Statement

In this work, we used generative AI tools for helping with editing the manuscript, implementing the baselines, and some assistance with code implementation for our method, all with author supervision and checking. We have not used generative AI tools for generating ideas or writing without human supervision. The first author verified LLM-generated research descriptions and LLM-generated code for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Xiaodong Cun and Chi-Man Pun. Split then refine: Stacked attention-guided resunets for blind single image visible watermark removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1184–1192, 2021.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition Workshops*, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps shadow removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 710–718, 2023a. doi: 10.1609/aaai.v37i1.25148.

Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14049–14058, 2023b.

Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8472–8480, 2024.

Tejas Jayashankar, Gary C. F. Lee, Alejandro Lancho, Amir Weiss, Yury Polyanskiy, and Gregory W. Wornell. Score-based source separation with applications to digital communication signals. In *Advances in Neural Information Processing Systems*, volume 36, pp. 5092–5125, 2023.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24174–24184, 2024.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022.

Alireza Kheirandish, Jihoon Hong, and Sara Fridovich-Keil. Klip: Localized distribution shift detection via kl-divergence with diffusion priors in inverse problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ronan Legin, Alexandre Adam, Yashar Hezaveh, and Laurence Perreault-Levasseur. Score-based likelihood characterization for inverse problems in the presence of non-gaussian noise. In *NeurIPS 2023 Workshop on Deep Learning and Inverse Problems*, 2023.

Jing Liang, Li Niu, Fengjun Guo, Teng Long, and Liqing Zhang. Visible watermark removal via self-calibrated localization and background refinement. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4426–4434, 2021.

Yang Liu, Zhen Zhu, and Xiang Bai. Wdnet: Watermark-decomposition network for visible watermark removal. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3685–3693, 2021.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20402–20411, 2023.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

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 *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

Giorgio Mariani, Irene Tallini, Emilian Postolache, Michele Mancusi, Luca Cosmo, and Emanuele Rodolà. Multi-source diffusion models for simultaneous music generation and separation. In *International Conference on Learning Representations*, 2024.

Zihao Mo, Junye Chen, Chaowei Fang, and Guanbin Li. Patchwiper: Leveraging dynamic patch-wise parameters for real-world visible watermark removal. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3537–3545. ACM, 2025. doi: 10.1145/3746027.3755107.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022.

Robin Scheibler, Youna Ji, Soo-Whan Chung, Jaeuk Byun, Soyeon Choe, and Min-Seok Choi. Diffusion-based generative speech source separation. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095310.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023.

Tristan S. W. Stevens, Hans van Gorp, Faik C. Meral, Junseob Shin, Jason Yu, Jean-Luc Robert, and Ruud J. G. van Sloun. Removing structured noise using diffusion models. *Transactions on Machine Learning Research*, 2025(3), 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=BvKYsaOVEn.

Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: Image processing in python. *PeerJ*, 2:e453, 2014. doi: 10.7717/peerj.453.

Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations*, 2023.

Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 650–656, 2022.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem: A discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8330–8339, 2021.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20895–20905, June 2025.

Xinyi Zhang, Naiqi Li, Jiawei Li, Tao Dai, Yong Jiang, and Shu-Tao Xia. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6782–6791, 2023.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1219–1229, 2023.

A Appendix

A.1 Experiment Details

Localized Structured Corruptions: All three benchmarks use the same 100 CelebA test images at 64×64 resolution. The natural-image corruptions are 32×32 pixel patches and CIFAR-10 are patches randomly chosen 16 to 48 pixels for the height and length. The corruptions follow the additive model $y = x + n$ on Ω, where x is the clean image and n is the spatially localized corruption. For SORD, we use $K = 32$ diffusion samples and a maximum of $M = 6$ refinement iterations. We use a noise schedule σ_m that starts at 0.4 and linearly decreases to 0.1.

We compare against DPS (Chung et al., 2023) baselines using the same CelebA clean-image diffusion prior. DPS-Inpaint treats the corrupted support as missing and reconstructs it by conditioning on the observations outside Ω. We additionally evaluate DPS under a Gaussian likelihood corruption model assumption within Ω, either using the oracle corruption mean $\mu = \mu_n$ or fixing $\mu = 0$. These variants test whether an explicit Gaussian noise model is sufficient to remove localized structured corruptions (it usually is not sufficient). We also compare with the two-prior SGM of Stevens et al. (2025), which jointly estimates the clean image and corruption using separate diffusion models; CelebA images with MNIST corruption is a main structured-noise setting in that work. For natural image-patch corruptions, we trained the SGM corruption model using using synthetic training examples generated by placing randomly selected image patches of size 26×26 to 38×38 on $64 * 64$ black background, separately for each type of image patch corruption. For the masked SGM variant, we enforce the known support by projecting all clean-image estimates at intermediate timesteps to match y on Ω^c.

Localized Audio Interference: Each example is a 10 second clip sampled at 22.05 kHz, and contains a single target source (piano, in Table 1) together with one interfering source. The interferer is restricted to a single contiguous window covering half of the clip, and this window is provided to all methods. We also evaluate the same three DPS configurations used for localized image corruptions. DPS uses only the target-source prior and the exact temporal window, with 250 Karras–Euler reverse steps and $y + \sigma_{\max}\epsilon$ initialization. DPS-Inpaint discards the measurements in the interference window, while the remaining variants use an in-window Gaussian noise assumption with either oracle $\mu = \mu_n$ or default $\mu = 0$. For SORD, we use the exact interference window and $K = 64$ diffusion samples.

Watermark Removal: CLWD is a paired synthetic benchmark containing $60,000$ training images generated from 160 colored watermark designs and $10,000$ test images generated from 40 additional watermark designs. The dataset provides clean background images and their corresponding watermarked observations, allowing the watermark support to be determined directly.

We consider three watermark-removal benchmarks. For FFHQ, we composite a subset of 12 CLWD train-split logos onto 25 held-out FFHQ faces at resolution 256×256; for Imagenette, we apply the same 12 logos to 25 Imagenette images at resolution 256×256. In both cases, the watermark support occupies approximately 3% of the image. For CLWD, we use its test split, containing 40 test-split logos and 99 test images at resolution 64×64, with an average corruption support of approximately 14% of the image. Composites use the per-pixel watermark opacity specified by the CLWD dataset.

The clean-image priors are off-the-shelf ADM (Dhariwal & Nichol, 2021) checkpoints for FFHQ and ImageNet-256 (for Imagenette). For CLWD, we use an EDM2 (Karras et al., 2024) ImageNet-64 model fine-tuned only on the clean images from the CLWD training split. For SLBR, we train a domain-matched SLBR model from scratch on paired data from the corresponding training domain: CLWD training pairs for CLWD, paired FFHQ examples for FFHQ, and paired Imagenette examples for Imagenette. Training uses all of the CLWD train/test watermarks except for the 12 watermarks which we used for testing. We have followed the published training code for reproducing results. PatchWiper uses the published PRWD (Mo et al., 2025) checkpoint in a blind, zero-shot setting and predicts its own watermark support. All DPS variants use the same clean prior and exact corruption support as SORD. DPS-Inpaint discards the observations within Ω, while the two likelihood-based variants use a per-image Gaussian noise assumption with mean set either to the oracle mean of the true corruption on Ω, $\mu = \mu_n$, or to zero as a standard default. We use 250 reverse sampling steps for resolution 64×64 and 1000 steps for resolution 256×256.

Table 3: **Ablation on the number of diffusion samples K.** Each entry reports corrupted-region PSNR (dB) of SORD as mean $\pm$ standard deviation over 25 images per benchmark. $K = 1$ is undefined because $\sigma^\star$ is the spread across samples. Best mean in each column is shown in **bold**.

K	Natural patch	MNIST digit	CIFAR-10 patch	CLWD watermark	ISTD shadow
2	18.26 $\pm$ 2.37	24.30 $\pm$ 2.50	17.44 $\pm$ 2.64	19.39 $\pm$ 3.72	27.45 $\pm$ 6.99
4	20.42 $\pm$ 2.28	25.46 $\pm$ 2.03	20.07 $\pm$ 2.96	20.63 $\pm$ 3.93	28.77 $\pm$ 6.06
8	21.74 $\pm$ 2.18	26.54 $\pm$ 1.81	21.36 $\pm$ 3.20	21.41 $\pm$ 3.81	29.60 $\pm$ 6.19
16	22.22 $\pm$ 2.08	27.09 $\pm$ 1.48	21.91 $\pm$ 3.54	21.68 $\pm$ 3.87	30.08 $\pm$ 5.99
32 (default)	22.52 $\pm$ 2.03	27.28 $\pm$ 1.43	22.23 $\pm$ 3.41	**22.04** $\pm$ 3.97	**30.13** $\pm$ 6.02
64	**22.54** $\pm$ 1.84	**27.40** $\pm$ 1.44	**22.41** $\pm$ 3.43	**22.04** $\pm$ 4.19	29.98 $\pm$ 6.20

Shadow Removal: For the ISTD benchmark, we use 100 real test shadows at resolution 64×64. To isolate restoration inside the declared shadow support, we recompose each example so that the corrupted input exactly matches the shadow-free target outside Ω; the shadow support occupies approximately 17% of the image on average. The clean prior is an EDM2 model fine-tuned only on shadow-free ISTD images. We additionally evaluate cross-domain transfer by applying 25 real ISTD shadow fields as multiplicative attenuation to held-out FFHQ and Imagenette images at resolution 256×256. These experiments use the corresponding off-the-shelf FFHQ and ImageNet-256 clean-image priors, respectively, with no shadow-specific training on either target domain.

A.2 Effect of the number of samples

The number of parallel samples K controls both the diversity of the initial inpainting hypotheses and the estimation of the ensemble spread $\sigma^\star$ used during refinement. We evaluate $K \in \{2, 4, 8, 16, 32, 64\}$ across the five image restoration settings in Table 3. Increasing K almost consistently improves the final reconstruction quality. Averaged over the five benchmarks, the corrupted-region PSNR of SORD increases from 21.37 dB at $K = 2$ to 24.84 dB at $K = 32$. The improvement largely saturates beyond this point. We therefore use $K = 32$ as the default throughout our experiments, providing a favorable trade-off between ensemble stability and computation.

A.3 Sampling Cost

We next examine whether the D_θ in SORD, currently instantiated using DDIM (Song et al., 2021), can be replaced by a faster sampling. We replace only the x-refinement step while keeping the inpainting initialization and n-refinement fixed. As shown in Table 4, reducing the number of steps does not significantly affect the final results.

A.4 Robustness to Imperfect Corruption Masks

Our formulation assumes access to the support Ω of the localized corruption. In practice, however, this support may be obtained from an approximate segmentation and may not coincide exactly with the true corrupted region. We therefore evaluate the sensitivity of SORD to systematic mask errors by perturbing the ground-truth support through dilation, erosion, and spatial shifts.

We evaluate the final iterate with $K = 32$ samples and report PSNR over the *true* corrupted region rather than over the perturbed mask. Table 5 reports mean $\pm$ standard deviation, with the intersection-over-union (IoU) between the supplied and true masks shown in parentheses.

Over-estimating the support is comparatively benign. Mask dilation produces a gradual degradation rather than a failure of the refinement procedure. In this regime, the primary consideration is that additional clean pixels are incorrectly treated as unknown and must be reconstructed by the diffusion prior.

Missing corrupted pixels is substantially more harmful. In contrast, under-estimating Ω can leave corrupted measurements outside the supplied support. Since SORD assumes that y_{Ω^c} is reliable

Table 4: **Number of DDIM steps in the x-refinement step.** Our main results use the "shipped" SORD variant, which runs the x-refinement step following RePaint (Lugmayr et al., 2022) over every ladder level at or below σ_m (13 levels at the top altitude). We test computationally cheaper variants that cap the ladder at 8 or 4 DDIM levels, with everything else unchanged (K-samples inpainting initialization). Image rows report corrupted-region PSNR (dB); audio rows report corruption-window SI-SDR (dB), both as mean $\pm$ standard deviation. Brackets report the number of network evaluations per image spent in the alternating loop (x-step and n-step) excluding the initialization, which is identical across columns. The highest mean in each row is shown in **bold**.

Bench	n	All levels (shipped)	8 DDIM levels	4 DDIM levels
Natural patch on CelebA	100	**21.96** $\pm$ 2.33 [1640]	**21.96** $\pm$ 2.33 [1352]	**21.96** $\pm$ 2.34 [840]
MNIST digit on CelebA	100	**27.16** $\pm$ 2.05 [1190]	**27.16** $\pm$ 2.05 [966]	27.15 $\pm$ 2.04 [614]
CIFAR-10 patch on CelebA	100	**22.05** $\pm$ 3.45 [1640]	**22.05** $\pm$ 3.45 [1352]	**22.05** $\pm$ 3.45 [840]
CLWD + watermark (declared α)	99	28.06 $\pm$ 6.48 [1184]	28.06 $\pm$ 6.48 [960]	**28.07** $\pm$ 6.47 [608]
PRWD + watermark (declared α)	100	30.75 $\pm$ 7.73 [1184]	30.75 $\pm$ 7.73 [960]	**30.76** $\pm$ 7.72 [608]
FFHQ + watermark (declared α)	25	**30.05** $\pm$ 3.99 [416]	**30.05** $\pm$ 3.99 [336]	30.03 $\pm$ 3.98 [208]
FFHQ + watermark, unseen logos	25	**26.76** $\pm$ 5.03 [416]	26.75 $\pm$ 5.02 [336]	26.74 $\pm$ 5.02 [208]
Imagenette + watermark (declared α)	25	**28.01** $\pm$ 7.15 [320]	**28.01** $\pm$ 7.15 [240]	27.98 $\pm$ 7.15 [144]
ISTD shadow (mult., gated)	100	**29.10** $\pm$ 5.55 [417]	29.07 $\pm$ 5.54 [257]	29.00 $\pm$ 5.52 [129]
ISTD shadow + FFHQ (mult., gated)	25	**20.34** $\pm$ 4.19 [418]	20.33 $\pm$ 4.19 [258]	20.32 $\pm$ 4.19 [130]
ISTD shadow + Imagenette (mult., gated)	25	**19.05** $\pm$ 6.50 [645]	**19.05** $\pm$ 6.50 [485]	18.48 $\pm$ 5.80 [212]
Slakh piano (SI-SDR, K=4, M=4)	100	**3.31** $\pm$ 6.75 [416]	3.22 $\pm$ 6.82 [288]	3.08 $\pm$ 6.87 [224]
Slakh guitar (SI-SDR, K=4, M=4)	50	**3.46** $\pm$ 5.14 [416]	3.27 $\pm$ 5.31 [288]	2.78 $\pm$ 5.93 [224]

Table 5: **Robustness to imperfect corruption masks.** We perturb the true corruption support using dilation, erosion, and spatial shifts and evaluate PSNR (dB) over the true corrupted region. Values in parentheses report IoU between the perturbed and true masks. The highest mean in each row is shown in **bold**.

Benchmark	Exact	Dilate 3 px	Dilate 5 px	Erode 1 px	Shift 2 px
Natural patch	**21.96** $\pm$ 2.33	19.94 $\pm$ 2.53 (0.73)	18.85 $\pm$ 2.54 (0.62)	13.84 $\pm$ 2.18 (0.88)	14.38 $\pm$ 2.08 (0.78)
MNIST digit	**27.16** $\pm$ 2.05	24.40 $\pm$ 2.72 (0.59)	23.32 $\pm$ 2.90 (0.47)	17.83 $\pm$ 1.48 (0.76)	15.20 $\pm$ 2.17 (0.60)
CLWD watermark	22.56 $\pm$ 4.33	19.60 $\pm$ 3.96 (0.59)	18.39 $\pm$ 3.72 (0.47)	**24.47** $\pm$ 4.91 (0.75)	22.14 $\pm$ 4.26 (0.61)
CIFAR-10 patch	**22.05** $\pm$ 3.45	19.53 $\pm$ 3.44 (0.69)	18.96 $\pm$ 3.47 (0.57)	12.11 $\pm$ 2.62 (0.85)	12.36 $\pm$ 2.70 (0.74)

and preserves those observations throughout refinement, such contamination is effectively locked into the reconstruction. Consequently, eroding the mask by only one pixel can cause a substantial decrease in mean PSNR.

These results indicate an asymmetric sensitivity to mask error: *over-covering the suspected corruption is substantially safer than leaving corrupted pixels outside the mask*.

An exception is watermark removal, as CLWD results are less sensitive to small mask erosions. Watermark boundaries often contain pixels with small or near-zero opacity, so removing these boundary pixels from the support has little effect. Accordingly, the one-pixel eroded mask performs comparably to, and in this experiment slightly above, the exact mask.

Qualitative examples are shown in Figure 4, illustrating that dilation mainly causes unnecessary regeneration of nearby clean content, whereas erosion or misalignment leaves visible residual corruption outside the region that SORD is permitted to modify.

Table 6: **Effect of removing the diffusion n-step across benchmarks.** The default method re-noises the estimated residual to noise level $\sigma^\star$ and denoises it with the clean-signal diffusion prior; "No DM n-step" instead uses the clipped raw residual while keeping all other components unchanged. Image results report PSNR (dB) over the corrupted support, while audio results report in-window SI-SDR (dB), as mean $\pm$ standard deviation. Audio experiments use $K = 4$ and $M = 4$ for computational efficiency. The highest mean in each column is shown in **bold**.

Method	Natural $n = 100$	MNIST $n = 100$	CIFAR-10 $n = 100$	CLWD WM $n = 99$	PRWD WM $n = 100$	FFHQ WM $n = 25$
DM n-step	**21.96** $\pm$ 2.33	**27.16** $\pm$ 2.05	**22.05** $\pm$ 3.45	**22.56** $\pm$ 4.33	**23.81** $\pm$ 5.17	26.68 $\pm$ 3.58
No DM n-step	20.16 $\pm$ 2.27	24.28 $\pm$ 2.36	20.40 $\pm$ 3.38	19.59 $\pm$ 4.61	22.81 $\pm$ 6.32	**29.22** $\pm$ 4.08

Method	ImageNet WM $n = 25$	ISTD $n = 100$	ISTD on FFHQ $n = 25$	ISTD on ImageNet $n = 25$	Slakh piano IV+OOV, $n = 100$	Slakh guitar IV+OOV, $n = 50$
DM n-step	24.92 $\pm$ 6.49	**29.10** $\pm$ 5.55	**20.34** $\pm$ 4.19	**19.05** $\pm$ 6.50	**3.31** $\pm$ 6.75	**3.46** $\pm$ 5.14
No DM n-step	**27.35** $\pm$ 7.41	28.83 $\pm$ 5.76	19.81 $\pm$ 3.97	17.23 $\pm$ 5.49	2.85 $\pm$ 7.31	2.78 $\pm$ 5.84

Table 7: **Effect of the re-noising schedule σ_m.** We vary the re-noising schedule used across refinement iterations while keeping all other components fixed. Results report PSNR (dB) over the corrupted support as mean $\pm$ standard deviation. Unless otherwise indicated, results are evaluated over 100 images. The highest mean in each column is shown in **bold**.

σ_m schedule	Natural	MNIST	CIFAR-10
$0.4 \to 0.1$, $M = 6$	**21.96** $\pm$ 2.33	**27.16** $\pm$ 2.05	**22.05** $\pm$ 3.45
Constant 0.8	20.50 $\pm$ 2.16	24.41 $\pm$ 1.92 ($n = 50$)	20.24 $\pm$ 3.56 ($n = 25$)
$\{0.8, 0.4, 0.2\}$	20.92 $\pm$ 2.37	26.21 $\pm$ 2.16	21.12 $\pm$ 3.49
$\{0.9, 0.6, 0.3, 0.1\}$	21.27 $\pm$ 2.38	26.78 $\pm$ 2.15	21.43 $\pm$ 3.48
$\{0.8, 0.4, 0.2, 0.1\}$	21.34 $\pm$ 2.39	26.89 $\pm$ 2.15	21.52 $\pm$ 3.51

Table 8: **Effect of support information and contaminant modeling.** We ablate the use of the exact corruption support and an additional contaminant diffusion model on the three localized CelebA benchmarks. Results report regional PSNR (dB) over Ω as mean $\pm$ standard deviation over 100 images. "No mask" treats the full image as the possible corruption support, whereas "Exact" uses the ground-truth support. SGM uses the joint ΠGDM sampler of Stevens et al. (2025). The best mean in each column is shown in **bold**.

Method	MNIST No mask	MNIST Exact	Natural No mask	Natural Exact	CIFAR-10 No mask	CIFAR-10 Exact
Input y	12.09 $\pm$ 2.49	12.09 $\pm$ 2.49	9.37 $\pm$ 1.63	9.37 $\pm$ 1.63	7.92 $\pm$ 1.91	7.92 $\pm$ 1.91
SORD	20.29 $\pm$ 3.38	27.16 $\pm$ 2.05	**16.42** $\pm$ 2.69	21.96 $\pm$ 2.33	**16.00** $\pm$ 3.38	22.05 $\pm$ 3.45
SORD + matched contaminant DM	18.44 $\pm$ 3.68	**29.12** $\pm$ 2.26	15.54 $\pm$ 2.51	**22.98** $\pm$ 2.15	14.91 $\pm$ 3.23	**22.08** $\pm$ 3.49
SGM / PiGDM, one prior	11.42 $\pm$ 3.11	15.60 $\pm$ 2.75	13.18 $\pm$ 2.04	14.77 $\pm$ 2.45	12.33 $\pm$ 2.95	13.98 $\pm$ 3.29
SGM / PiGDM, two priors	**25.78** $\pm$ 3.42	22.51 $\pm$ 3.89	11.45 $\pm$ 2.11	16.96 $\pm$ 3.32	11.66 $\pm$ 2.83	11.25 $\pm$ 2.77

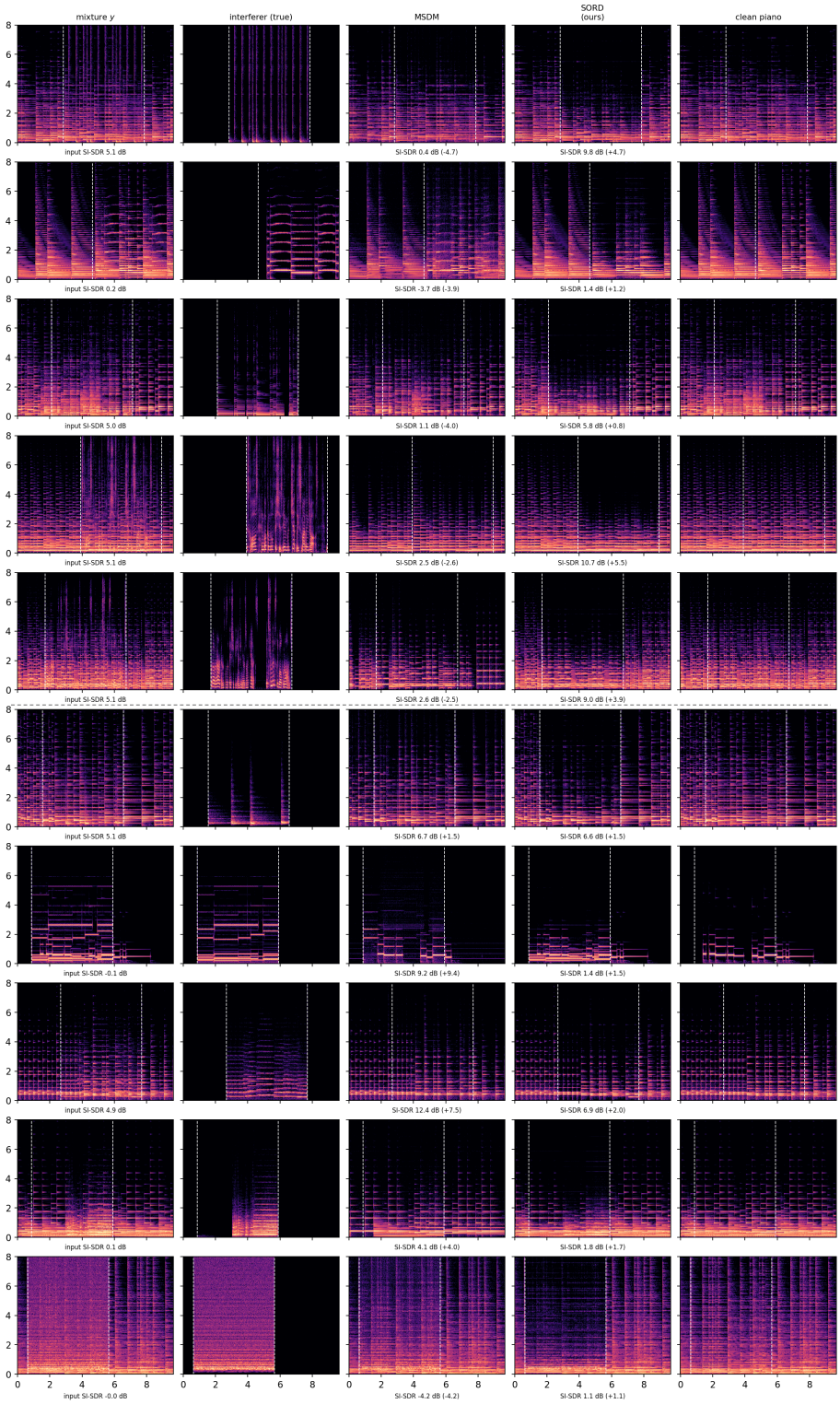

Figure 3: **Visualization of localized audio interference removal.** Each row shows a Slakh piano clip corrupted by a temporally localized interferer within the dashed interval. Columns show the mixture y, the ground-truth interferer, MSDM, SORD using only the clean-piano prior, and the clean target. Rows 1–5 use IV interferers (drums, guitar, bass, and speech), for which MSDM has corresponding source priors, while Rows 6–10 use OOV interferers (strings, organ, reed, brass, and synthesizer pad), for which no matched MSDM prior is available. Values report in-window SI-SDR and, in parentheses, improvement over the input mixture.



Figure 4: **Qualitative effect of corruption-mask errors.** We compare SORD using the exact corruption support or with dilated, eroded, or shifted masks. Over-estimating the support causes the clean prior to regenerate some originally uncorrupted pixels, whereas under-estimating the support leaves portions of the corruption outside Ω, where they are treated as trusted observations and therefore are not corrected.

Table 9: **Localized structured corruption removal across image and audio domains.** Audio results report windowed SI-SDR improvement (SI-SDRi, dB) for piano, guitar, bass, and drums targets under localized instrumental and speech interference. All values are mean ± standard deviation. IV and OOV denote MSDM in-vocabulary and out-of-vocabulary interferers, respectively; for speech, IV uses MSDM augmented with a speech prior, while OOV uses the original four-source MSDM without a speech prior. The setting $g = 1$ denotes the unnormalized mixture. SGM and MSDM use explicit contaminant/source priors, whereas SORD uses only the clean target prior. The highest mean in each row is shown in **bold**.

Audio Interferer	Mix	n	DPS Inpaint	DPS $\mu = \mu_n$	DPS $\mu = 0$	MSDM	SORD
Piano target							
IV instrument	SIR 0	60	−0.34 ± 6.17	−0.01 ± 0.05	−0.06 ± 0.43	**5.32** ± 7.64	1.12 ± 5.11
IV instrument	SIR 5	60	−0.12 ± 5.67	0.00 ± 0.07	0.01 ± 0.09	1.40 ± 6.85	**2.74** ± 6.23
IV instrument	$g = 1$	60	0.34 ± 0.92	−0.02 ± 0.13	−0.02 ± 0.12	**6.48** ± 9.06	1.29 ± 2.58
Speech (IV)	SIR 0	160	1.02 ± 1.18	−0.01 ± 0.10	0.00 ± 0.05	**8.25** ± 8.64	5.51 ± 3.37
Speech (IV)	SIR 5	160	1.85 ± 2.02	−0.03 ± 0.28	0.00 ± 0.05	2.00 ± 10.03	**6.40** ± 3.39
OOV instrument	SIR 0	160	−0.43 ± 17.01	−1.14 ± 14.35	−1.22 ± 15.50	−0.24 ± 16.94	**1.15** ± 17.54
OOV instrument	SIR 5	160	0.03 ± 17.15	−1.21 ± 15.50	−1.13 ± 14.35	−2.67 ± 18.39	**1.97** ± 17.54
OOV instrument	$g = 1$	160	−0.34 ± 17.08	−1.23 ± 15.50	−1.23 ± 15.50	−1.54 ± 17.40	**0.79** ± 17.41
Speech (OOV)	SIR 0	160	1.02 ± 1.18	−0.01 ± 0.10	0.00 ± 0.05	3.56 ± 5.83	**5.51** ± 3.37
Speech (OOV)	SIR 5	160	1.85 ± 2.02	−0.03 ± 0.28	0.00 ± 0.05	0.11 ± 6.81	**6.40** ± 3.39
Guitar target							
IV instrument	SIR 0	60	0.52 ± 0.85	0.00 ± 0.03	0.00 ± 0.03	**4.83** ± 8.49	2.24 ± 3.02
IV instrument	SIR 5	60	1.18 ± 1.69	0.01 ± 0.10	0.00 ± 0.09	−0.04 ± 8.97	**4.34** ± 3.95
IV instrument	$g = 1$	60	0.43 ± 0.74	0.00 ± 0.03	0.01 ± 0.03	**5.36** ± 8.26	1.29 ± 2.68
Speech (IV)	SIR 0	110	0.62 ± 4.81	−0.21 ± 2.26	−0.22 ± 2.26	3.86 ± 11.14	**4.73** ± 3.83
Speech (IV)	SIR 5	110	1.70 ± 5.29	−0.26 ± 2.74	−0.26 ± 2.75	−1.72 ± 14.26	**6.14** ± 4.83
OOV instrument	SIR 0	160	−2.66 ± 29.39	−3.53 ± 25.66	−3.53 ± 25.68	−3.92 ± 31.61	**−0.74** ± 30.23
OOV instrument	SIR 5	160	−1.76 ± 29.59	−3.52 ± 25.68	−3.53 ± 25.68	−4.47 ± 30.81	**−0.01** ± 30.19
OOV instrument	$g = 1$	160	−3.39 ± 29.93	−3.89 ± 26.13	−3.86 ± 25.94	−5.88 ± 32.06	**−2.03** ± 30.70
Speech (OOV)	SIR 0	110	0.62 ± 4.81	−0.21 ± 2.26	−0.22 ± 2.26	−0.78 ± 9.12	**4.73** ± 3.83
Speech (OOV)	SIR 5	110	1.70 ± 5.29	−0.26 ± 2.74	−0.26 ± 2.75	−1.05 ± 7.60	**6.14** ± 4.83
Bass target							
IV instrument	SIR 0	60	3.92 ± 3.57	0.00 ± 0.02	0.00 ± 0.03	**21.80** ± 8.59	9.83 ± 8.02
IV instrument	SIR 5	60	6.80 ± 5.88	0.00 ± 0.07	0.00 ± 0.07	**19.26** ± 7.81	12.31 ± 5.96
IV instrument	$g = 1$	60	4.64 ± 4.86	0.00 ± 0.03	0.00 ± 0.03	**21.54** ± 7.98	10.06 ± 7.64
Speech (IV)	SIR 0	110	4.91 ± 2.99	0.01 ± 0.02	0.01 ± 0.01	8.69 ± 2.48	**12.93** ± 5.69
Speech (IV)	SIR 5	110	7.85 ± 4.40	0.00 ± 0.03	0.00 ± 0.03	4.03 ± 2.59	**15.27** ± 3.80
OOV instrument	SIR 0	160	7.45 ± 5.75	0.01 ± 0.08	0.01 ± 0.09	**18.58** ± 9.82	14.32 ± 7.84
OOV instrument	SIR 5	160	10.53 ± 6.14	0.00 ± 0.05	0.00 ± 0.07	**16.43** ± 8.48	14.24 ± 5.19
OOV instrument	$g = 1$	160	8.51 ± 6.15	0.02 ± 0.19	0.00 ± 0.10	**17.52** ± 8.83	13.51 ± 6.92
Speech (OOV)	SIR 0	110	4.91 ± 2.99	0.01 ± 0.02	0.01 ± 0.01	**17.74** ± 6.67	12.93 ± 5.69
Speech (OOV)	SIR 5	110	7.85 ± 4.40	0.00 ± 0.03	0.00 ± 0.03	**15.88** ± 6.58	15.27 ± 3.80
Drums target							
IV instrument	SIR 0	60	1.25 ± 1.00	0.01 ± 0.09	0.02 ± 0.10	**19.88** ± 6.54	9.31 ± 5.48
IV instrument	SIR 5	60	1.85 ± 1.36	−0.01 ± 0.13	0.00 ± 0.12	**16.54** ± 7.10	9.62 ± 4.22
IV instrument	$g = 1$	60	1.28 ± 1.09	0.00 ± 0.12	0.01 ± 0.10	**19.67** ± 6.79	9.47 ± 5.55
Speech (IV)	SIR 0	110	1.32 ± 0.55	0.03 ± 0.08	0.02 ± 0.07	**11.85** ± 6.41	8.89 ± 3.31
Speech (IV)	SIR 5	110	1.48 ± 0.67	0.03 ± 0.10	0.01 ± 0.10	7.56 ± 8.22	**10.05** ± 3.48
OOV instrument	SIR 0	160	−0.54 ± 23.97	−2.38 ± 21.43	−2.37 ± 21.32	**12.60** ± 25.89	9.24 ± 25.33
OOV instrument	SIR 5	160	0.02 ± 24.04	−2.36 ± 21.10	−2.39 ± 21.17	**9.79** ± 25.90	8.22 ± 25.12
OOV instrument	$g = 1$	160	−0.96 ± 24.32	−2.56 ± 21.62	−2.59 ± 21.34	**10.58** ± 26.70	8.26 ± 25.77
Speech (OOV)	SIR 0	110	1.32 ± 0.55	0.03 ± 0.08	0.02 ± 0.07	**9.93** ± 4.36	8.89 ± 3.31
Speech (OOV)	SIR 5	110	1.48 ± 0.67	0.03 ± 0.10	0.01 ± 0.10	8.89 ± 4.98	**10.05** ± 3.48