This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

KLIP: localized distribution shift detection via KL-divergence with diffusion priors in Inverse Problems

Alireza Kheirandish* Jihoon Hong* Sara Fridovich-Keil
Georgia Institute of Technology
School of Electrical and Computer Engineering
{akheirandish3, jhong392, sfk}@gatech.edu

Abstract

Diffusion models have shown promising performance as data-driven priors for computational imaging, as well as some capacity to detect out-of-distribution (OOD) images. However, existing approaches to OOD detection often require some knowledge of the shifted distribution, fail to detect subtle or localized distribution shifts, and operate on full images, rather than the indirect measurements available in inverse problems. We propose an OOD detection metric based on the Kullback-Leibler divergence between the diffusion prior and the posterior distribution, that (i) does not require any calibration data or knowledge of the shifted distribution, and (ii) can detect whole images as OOD as well as localize OOD patches within an image. Experimentally, we show that this metric can detect subtle yet semantically meaningful distribution shifts, such as the shift from healthy liver CT scans to those with tumors, and generalizes across different types of diffusion models, datasets, and inverse problems. Our code can be found at https://github.com/voilalab/KLIP.

1. Introduction

Distribution shift detection is a foundational problem in machine learning and computational imaging, especially when models are to be deployed in an open world. In many applications, the distribution shifts that are most important to catch are also some of the most difficult to detect, often involving subtle or localized changes in an image. In medical imaging, for example, out-of-distribution (OOD) features may consist of small lesions, tumors, or tears that are essential to the diagnostic value of the image. We propose to use diffusion models trained on in-distribution (ID) images to simultaneously perform underdetermined image reconstruction and detection of spatially localized OOD features.

Many approaches have been proposed for OOD detection using diffusion models. The most statistically rigorous OOD detection metrics are often based on conformal prediction [3, 4, 20], but these require access to a calibration dataset of OOD images, which is impossible to acquire for the long tail of rare abnormalities that appear in open-world images. Other approaches to OOD detection are more heuristic, but still require an OOD calibration dataset [25], require training multiple diffusion models [32], operate on full images rather than indirect measurements available in computational imaging [10], or focus on distinguishing OOD images with global rather than local OOD features [33]. We instead propose an OOD detection metric that (i) does not require any knowledge or samples of OOD images,

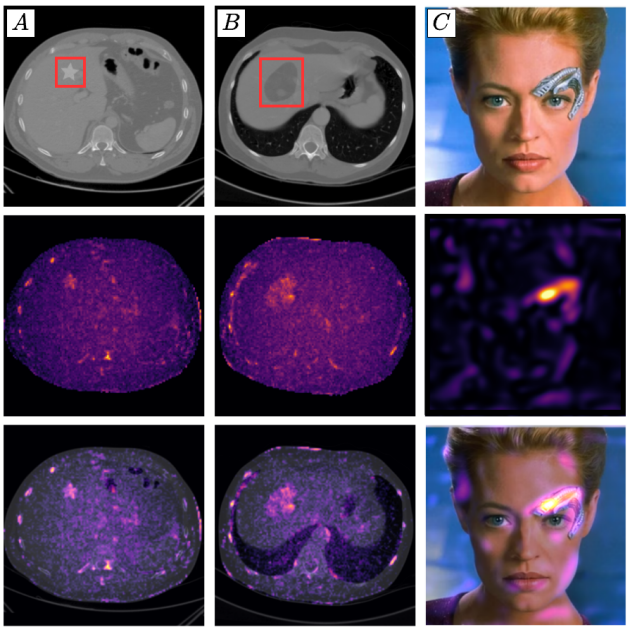

Figure 1. KLIP (*middle row*) is able to identify the precise locations of local distribution shifts from diffusion model samples acquired through posterior sampling in an inverse problem (here, underdetermined CT and Gaussian deblurring). The diffusion model is trained on only ID images (healthy CT scans and celebrity faces), and KLIP has no access to other OOD examples.

*Equal contribution. Order of names was decided by a coin toss.

(ii) can be applied to indirect measurements of an image that are available at inference time in an inverse problem, and (iii) can detect small and localized, yet semantically meaningful, distribution shifts.

Our core idea is that the Kullback-Leibler (KL) divergence between the prior distribution of the image, and the posterior distribution of the image given the measurements, should be larger for OOD images and larger for OOD regions within an OOD image. This KL divergence can be estimated using only the update steps generated throughout diffusion sampling conditioned on a set of measurements, requiring no OOD calibration data. Using the knowledge that diffusion models generate different components of an image at different stages of the sampling process [8, 30], we further refine this KL divergence metric by estimating its restriction over specific time windows of the sampling trajectory, and over spatially localized blocks in the image.

We evaluate our proposed KL divergence based OOD detection metric, KLIP, on a range of image reconstruction tasks with localized OOD features. Our primary results focus on sparse view X-ray computed tomography (CT), where the diffusion prior is trained over a dataset of healthy scans. We demonstrate dataset-level and image-level OOD detection of both synthetic stars and realistic synthetic liver tumors [15], across two different diffusion model backbones, one trained on whole images [37] and another trained on image patches [14]. We also demonstrate KLIP's generalization potential via image-level detection of anomalous character facial features in a Gaussian de-blurring task, using a diffusion prior trained on celebrity faces [41]. As shown in Figure 1, KLIP can locate the positions of (a) a synthetic star artifact and (b) tumor within a liver CT scan, and (c) the implant around the left eye of Seven of Nine, a character from *Star Trek: Voyager*. To summarize, we:

- Introduce a near-OOD detection metric, KLIP, based on the timestep and block restricted KL divergence between prior and posterior distributions in an inverse problem. We find that these timestep and block restrictions offer a rich window into the sampling process under distribution shift. KLIP is calibration-free and can be calculated using a single diffusion model trained on ID images.
- Demonstrate that KLIP can detect dataset-level and image-level distribution shift, when the OOD features are subtle and spatially localized. KLIP generalizes across different inlier distributions (CT scans and human faces), different localized OOD features (synthetic stars, realistic liver tumors, scars, and makeup), and different diffusion model architectures (whole-image and patch-based).

2. Related Works

Many methods have been proposed to leverage diffusion models for distribution shift detection.

Calibration-based methods. Many approaches to OOD detection, with and without diffusion models, rely on a calibration dataset of OOD images with which to estimate an ID–OOD decision boundary [25] or train a separate OOD diffusion model [32]. These methods include conformal prediction [4, 13, 39], which can provide statistically rigorous confidence intervals around OOD image reconstructions. They also include methods that quantify uncertainty at the pixel or region level within an image [3, 20], and adaptive schemes to collect additional measurements to reduce uncertainty [16, 44]. While these methods can work well when both ID and OOD data is available, they are not applicable to distribution shift detection when it first arises, or to detecting the long tail of rare OOD images that exist in an open world.

Uncertainty quantification methods. Many approaches have been proposed for uncertainty quantification, but are not specifically focused on OOD detection or uncertainty arising from distribution shifts. For example, Monte Carlo methods have been proposed to estimate uncertainty due to limited measurements [1, 2, 29], and ensemble methods have been proposed to quantify uncertainty over the weights of the learned prior [17, 21]. Some approaches to uncertainty quantification avoid the need for a calibration dataset by simulating calibration data with a physical model [38], especially in inverse problems with strong physical constraints such as partial differential equations [29, 45, 50]. Our focus is instead on calibration-free detection of local OOD features during image reconstruction, in generic linear inverse problems with diffusion priors. Existing approaches to calibration-free detection of localized OOD features have focused on detecting anomalies in otherwise highly consistent images, such as those for quality control in manufacturing [11], rather than in more diverse ID images.

OOD detection with generative priors. Methods have also been proposed within our focus area of calibration-free OOD detection with generative (specifically diffusion-based) priors, though these generally do not allow for detection of localized OOD features in images that are otherwise in-distribution [10, 33, 40]. A common assumption of OOD detection methods is that OOD samples correspond to regions of low density under the prior distribution [6]; [26] showed that this is not the case for generative models, and [27] proposed an alternative OOD detection test based on the notion of a distribution's *typical set*. While this approach can improve dataset-level OOD detection, it has not been evaluated for capacity to detect the sort of highly localized OOD features that are our focus.

3. Preliminaries

We begin by introducing the diffusion models and posterior sampling methods that are the context for KLIP.

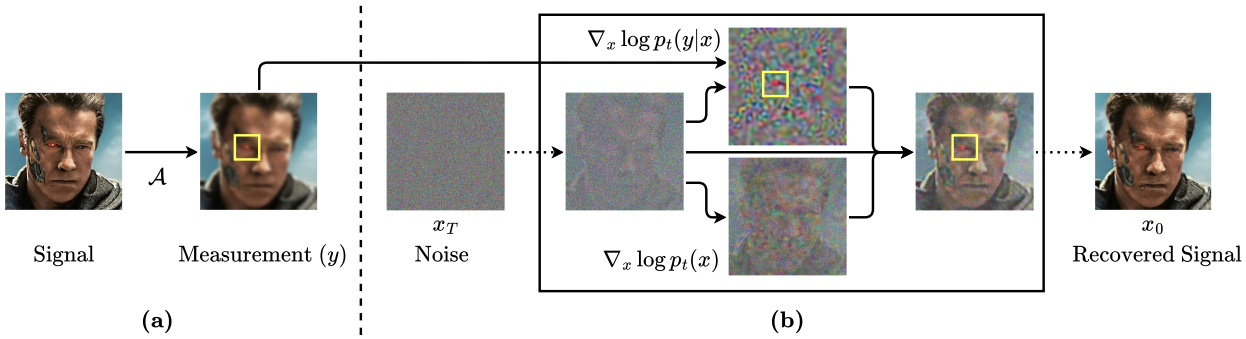

Figure 2. **Solving inverse problems using posterior sampling with a diffusion model**, for a Gaussian de-blurring example. (a) Forward model of gaussian blur. (b) Posterior sampling using a diffusion model to recover the signal from the measurement y following Equation (6). We can see from the yellow boxes highlighting the Terminator's red eye that the score of the likelihood guides sampling to be consistent with y, especially where there is a locally OOD feature.

Score-Based Diffusion Models. To sample from a data distribution $p(x)$, which is often high dimensional and multi-modal, score-based diffusion models [7, 35] first define two processes: forward and backward. The forward process is described by a stochastic differential equation (SDE) defined over time $t \in [0, T]$:

$$dx = f(x,t)dt + g(t)dw, \tag{1}$$

that we call the forward SDE. Here $f(x,t)$ is the drift coefficient, $g(t)$ is the diffusion coefficient, and w is a standard Brownian motion. We denote the marginal distribution of x at time t as $p_t(x)$, and $p_0(x)$ is by definition equal to the data distribution $p(x)$. The coefficients of the SDE are determined such that $p_T(x)$ approximates a tractable distribution that we can easily sample from, typically $\mathcal{N}(0, \sigma_T^2 I)$.

The backward process is described by another SDE defined from $t = T$ to $t = 0$, also called the reverse SDE of Equation (1), as its marginal distribution at each time coincides with $p_t(x)$. It takes the form:

$$dx = \big(f(x,t) - g(t)^2 \nabla_x \log p_t(x)\big)dt + g(t)d\overline{w}, \tag{2}$$

where $f(x,t)$ and $g(t)$ are precisely the coefficients of Equation (1), $\nabla_x \log p_t(x)$ is the score of the marginal distribution $p_t(x)$ of the forward SDE at time t, and $\overline{w}$ is the reverse-time standard Brownian motion.

With the forward and backward processes defined as above, we can easily sample from the data distribution: first sample x_T from $p_T(x)$, and simulate the reverse SDE in Equation (2) until we reach $x_0 \sim p_0(x) = p(x)$. The sampling can be performed in various ways, including Euler-Maruyama discretization of the reverse SDE:

$$\Delta x_t = \big(f(x,t) - g(t)^2 s_\theta(x_t; t)\big)\Delta t + g(t)z, \tag{3}$$

where $z \sim \mathcal{N}(0, I)$. Here, since $\nabla_x \log p_t(x)$ is not tractable, a diffusion model $s_\theta(x; t)$ trained with a score matching loss is used to approximate it during sampling.

Inverse Problems. Diffusion models are widely used as generative priors to solve inverse problems [9, 37, 42, 46, 47] of the following standard form:

$$y = \mathcal{A}(x) + \epsilon, \tag{4}$$

where x is an image to be reconstructed, $\mathcal{A}$ models the linear [9, 19] or nonlinear [42, 46, 47] forward process of collecting measurements y, and ϵ is measurement noise. These works aim to sample from the posterior distribution $p(x|y)$, using an unconditional diffusion model as well as a measurement y of the ground truth image x acquired through the measurement process in Equation (4). Since the diffusion model is trained to sample from the prior distribution $p(x)$, the sampling process is modified to ensure consistency with the measurements y by replacing the score of the prior distribution $p(x)$ with the score of the posterior $p(x|y)$:

$$\nabla_x \log p_t(x|y) = \nabla_x \log p_t(x) + \nabla_x \log p_t(y|x). \tag{5}$$

Here we use Bayes' rule to decompose the posterior score $\nabla_x \log p_t(x|y)$ into the sum of the prior score $\nabla_x \log p_t(x)$ and the likelihood score $\nabla_x \log p_t(y|x)$, which encodes consistency with the measurements y.

While the prior score can be estimated using a diffusion model trained on $p(x)$, the likelihood score is typically intractable. Therefore, previous works such as Diffusion Posterior Sampling (DPS) [9] and Song et al. [37] have proposed approximations or proxies of the likelihood score with a function $s_l(x_t, y; t)$, which we denote with the subscript l for likelihood, and augment the unconditional sampling process as:

$$\Delta x_t = \Big(f(x,t) - g(t)^2\big(s_\theta(x_t;t) + s_l(x_t, y; t)\big)\Big)\Delta t + g(t)z. \tag{6}$$

DPS approximates the likelihood score directly under assumptions on the measurement noise ϵ, and rigorously proved bounds on the approximation error. Meanwhile,

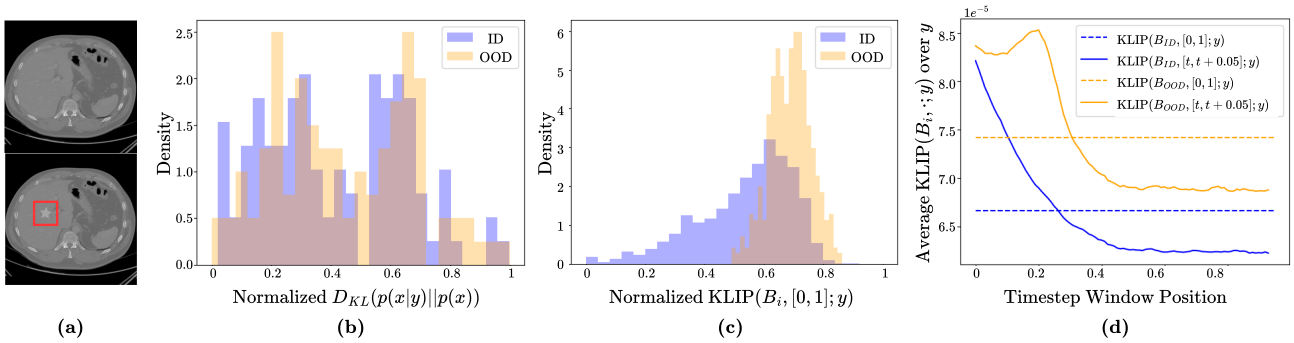

Figure 3. **Motivation for block and timestep restriction:** (a) An example ID image of a healthy liver CT scan (*top*) and corresponding OOD image with a localized synthetic star artifact (*bottom*). (b) Histogram of the prior–posterior KL divergence (Equation (11)) for ID images and OOD images, normalized to $[0, 1]$. The ID and OOD images are not well separated. (c) Histogram of the block restricted KL divergence over the same dataset, normalized to $[0, 1]$. OOD blocks are those that contain the synthetic star artifact. These distributions are much better separated, though there is still some overlap. (d) Average of block restricted KL divergence over the entire sampling trajectory (dotted lines) and over a sliding timestep window (solid line). Further restricting the blockwise KL divergence to focus on a specific time window can enhance separation between ID and OOD blocks.

Song et al. [37] proposed to insert an additional proximal optimization step at each time t during sampling, and showed experimentally that adding this step is sufficient to sample from $p(x|y)$ instead of $p(x)$ for a limited class of linear forward models. Therefore, we consider the updates from the optimization step as a proxy for the difference between Equation (3) and Equation (6), which is the likelihood score scaled by $g(t)^2$.

4. Methods

Prior–Posterior KL Divergence. We posit that the KL divergence between the prior distribution $p(x)$ and the posterior distribution $p(x|y)$, denoted as $D_{KL}(p(x|y)\|p(x))$, is an effective tool for both dataset-level and image-level distribution shift detection. Intuitively, the more the measurement y "pulls" the posterior away from the prior, the higher the chance that x is OOD. As a more concrete motivation, consider a toy example of a Gaussian prior $p(x) = \mathcal{N}(0, \sigma_1^2)$, with a forward model $y = x + \epsilon$ under the noise model $\epsilon \sim \mathcal{N}(0, \sigma_2^2)$. The posterior in this example is also a Gaussian, with mean proportional to the measurement y:

$$p(x|y) = \mathcal{N}\Big(\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} y, \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}\Big). \tag{7}$$

The KL-divergence between the prior and the posterior is

$$D_{KL}\big(p(x|y)||p(x)\big) = \frac{\sigma_1^2 y^2}{2(\sigma_1^2 + \sigma_2^2)^2} + \text{constant}. \tag{8}$$

This KL divergence grows quadratically with $|y|$. Under the forward model in this example, a signal $x^\star$ that gives rise to an observation y with a larger magnitude is increasingly unlikely under the prior, so the KL-divergence effectively captures where $x^\star$ lies relative to the prior distribution.

In general, the KL divergence between the prior and posterior distributions is not always easily accessible. However, under certain regularity conditions [36], the KL divergence of two arbitrary distributions $D_{KL}(p\|q)$ can be written in terms of the scores of the marginal distributions p_t and q_t, under a fixed SDE of the form in Equation (1). Specifically,

$$D_{KL}(p\|q) = \frac{1}{2} \int_0^T E_{x \sim p_t} \Big[\|g(t) h(x, t)\|_2^2 \Big] dt, \tag{9}$$

where $p_0 = p$, $q_0 = q$, $g(t)$ is the drift of the SDE, and

$$h(x, t) = \nabla_x \log p_t(x) - \nabla_x \log q_t(x). \tag{10}$$

While Equation (9) also cannot be computed in general unless the two scores in Equation (10) are available, we observe that posterior sampling offers access to estimates of these scores for the prior and the posterior. Replacing p and q with the posterior distribution $p(x|y)$ and the unconditional training data distribution $p(x)$, respectively, following Equation (5), $h(x, t)$ becomes the score of the likelihood $\nabla_x \log p_t(y|x)$. As described in Section 3, posterior sampling estimates this score with $s_l(x, y; t)$, which yields the following approximation of the KL divergence:

$$D_{KL}\big(p(x|y)||p(x)\big) = \frac{1}{2} \int_0^T E_{x \sim p_t(x|y)} \Big[\|g(t) s_l(x, y; t)\|_2^2 \Big] dt. \tag{11}$$

To compute this KL divergence given measurement y, we first generate multiple samples following the discrete posterior sampling process described in Equation (6), each with different random $z \in \mathcal{N}(0, I)$. Along the way, we collect the estimated likelihood scores $s_l(x_t, y; t)$ corresponding to each noise z and each timestep t. Afterwards, we average

the norm in Equation (11) over random samples to approximate the expectation, and sum over timesteps, to approximate the integral in Equation (11).

KLIP: Block & Timestep Restriction. To effectively detect and identify spatially localized OOD image features, we consider the restriction of the Kullback-Leibler (KL) divergence in both space and time.

First, we divide each score $s_l(x, y; t) \in \mathbb{R}^{D \times D}$ into a total of N_B blocks, each with size $D_B \times D_B$. We denote the restriction of the score to each block B_i as $s_l(x, y; t)\big|_{B_i}$. This is motivated by a longstanding statistics and information-theory literature on non-parameteric estimation of the KL divergence using histogram or data-dependent partitions, which represents the divergence as an aggregation of cell-wise contributions [28, 34, 43]. Since Equation (11) uses a squared l_2 norm over spatial coordinates, restricting to blocks yields a localized contribution. This helps mitigate signal dilution when the OOD feature occupies a small region, by preventing the local distribution shift from being averaged out by the rest of the image, as shown in Figure 3. Analogous to the histogram bin widths in previous works, block size D_B controls the trade-off between better localization and higher variance.

Second, we further restrict the timestep ranges to integrate over in Equation (11). For each block B_i, we define KLIP$(B_i, [t_0, t_1]; y)$ by restricting Equation (11) to that spatial block as well as the timestep range $[t_0, t_1]$:

$$\text{KLIP}(B_i, [t_0, t_1]; y) = \frac{1}{2} \int_{t_0}^{t_1} E_{x \sim p_t(x|y)} \left[\| g(t) s_l(x, y; t)\big|_{B_i} \|_2^2 \right] dt, \quad (12)$$

where $t_0 < t_1$. This time-range restriction is inspired by previous works [8, 30, 48] studying the different roles each timestep plays during sampling, including generating features at different scales [30, 48] and modeling images over a lower-dimensional manifold at certain timesteps [8].

The benefits of these restrictions are demonstrated in Figure 3. The distributions of whole image prior–posterior KL divergence for ID images (normal CT scans) and OOD images (CT scans with small synthetic star artifacts), shown in Figure 3 (b), do not display any separation. In contrast, the distributions of block-restricted KL divergence for blocks containing and not containing the OOD star artifact, shown in Figure 3 (c), are much better separated. Similarly, timestep restriction can reveal patterns that are informative for distribution shift detection. Figure 3 (d) shows that when integrated over a sliding timestep window of size 0.05 (relative to $T = 1$), OOD blocks exhibit a bump at around $t = 0.3$, that is absent in ID blocks (see solid lines). This pattern is not captured when integrating over the entire timestep range $[0, 1]$ (see dotted lines).

5. Evaluation

Models & Datasets. We evaluate the effectiveness of our proposed OOD metric KLIP (Equation (12)), across different settings. We first consider computed tomography (CT) as our inverse problem, for which we consider two distinct diffusion models: a predictor-corrector based model [35, 37] and a patch-based model (PaDIS) [14]. We selected these models to assess the generalizability of KLIP, as these models differ not only in the sampling algorithm, but also in how the likelihood score is approximated. Further details on the approximations involved in each method for posterior sampling can be found in Section 7.

We define the inlier distribution as healthy abdominal CT scans, and liver tumors as the local OOD features absent in the inlier distribution. We train both diffusion models on the training set of the Combined Healthy Abdominal Organ Segmentation (CHAOS) dataset [18], which contains 512 $\times$ 512 resolution abdominal CT scans of 40 different patients. Since the patients whose CT scans are included in the dataset are potential liver donors with healthy livers, liver tumors are guaranteed to be both localized and out of distribution. We select a subset of 100 images from the test set as the *ID set*, and generate an *OOD set* by adding realistic synthetic tumors to each image of the *ID set*. We generate synthetic liver tumors following [15], whose results closely resemble the tumors found in real patients.

We also validate the localization of OOD features on another inverse problem, Gaussian de-blurring of face images. For this task, we use CelebA [23] as the inlier distribution, and perform Gaussian de-blurring with a DDPM-based [12] diffusion model [41] trained on that dataset. We generate two locally OOD datasets: one by adding realistic synthetic scars of various shapes to the faces of 20 different CelebA test images, and another by collecting 15 images of human-like movie and television characters, including those from Star Trek and Harry Potter, whose faces contain localized unusual features not present in the ID CelebA dataset [23].

Experiments. We first evaluate the performance of KLIP on distinguishing OOD and ID CT scans, based on their projection measurements. For each image in the *ID set* and the *OOD set*, we simulate sparse-view CT measurements (with 24 projection angles) and compute KLIP, approximating the expectation in Equation (12) with the average over 5 samples for each measurement. For image-level OOD detection we use only the images in the *OOD set*, and mask out the parts of CT scan outside the body mass. We repeat the image-level experiment with CelebA [23] as inlier distribution by using KLIP to localize the OOD features on human and human-like faces. On this dataset, the expectation is estimated over 8 samples instead of 5. In Section 12 we report an ablation study over the number of samples used to approximate the expectation.

Table 1. **AUC results for dataset-level and image-level OOD detection**, for different OOD artifacts, models, and inverse problems. We mark with † those AUC values that correspond to settings for which we tuned the KLIP hyperparameters. Other experiments use the same set of hyperparameters without further refinement. The best AUC for each task and dataset is underlined. Note that CutPaste and SimpleNet operate directly on images rather than measurements, making their OOD detection task somewhat easier.

					Predictor-Corrector [37]						**PaDIS [14]**			**DDPM [41]**		
	Dataset (CT)						**Image (CT)**							**Image (Deblur)**		
	CutPaste	SimpleNet	NLL	DiffPath	D_{KL}	KLIP	D_{KL}	KLIP	CutPaste	SimpleNet	D_{KL}	KLIP		D_{KL}	KLIP
Star	0.999	0.993	0.586	0.688	0.541	0.855†	0.837	0.912†	0.978	0.999	0.841	0.889	**Scar**	0.675	0.867
Tumor	0.505	0.504	0.535	0.368	0.602	0.776	0.856	0.878	0.441	0.592	0.672	0.732	**Character**	0.482	0.772

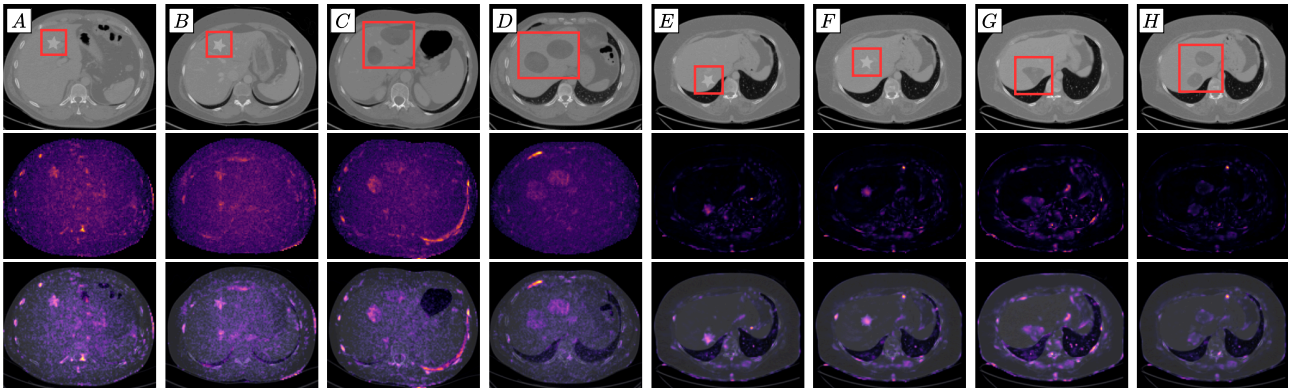

Figure 4. **Visualization of image-level OOD detection in sparse-view CT scans.** *Top:* Images in the *Tuning set* and *OOD set*. Red boxes annotate the stars and tumors. *Middle:* Heatmaps of KLIP computed pixel-wise (i.e. $D_B = 1$). *Bottom:* Heatmaps overlaid on images, to show localization. Columns A through D use a predictor-corrector based diffusion model, and columns E through H use a patch-based diffusion model. For each model, we show two examples from the *Tuning set* (stars) and two from the *OOD set* (tumors).

We note that KLIP has multiple hyperparameters: the block size D_B and the time ranges $[t_0, t_1]$. To choose these values, we generate a *Tuning set* by adding synthetic star-shaped artifacts to each CT scan image in the *ID set*. Then, we perform a grid search over the hyperparameters to maximize OOD detection performance with respect to this *Tuning set* using the predictor-corrector based model [37]. We tune hyperparameters separately for each task (dataset-level and image-level OOD detection) but use the same hyperparameter values for experiments on the more realistic tasks of tumor detection in sparse-view CT and face artifact detection in Gaussian de-blurring, for all models.

Baselines & Metrics. We compare KLIP against existing OOD detection techniques, across both dataset-level and image-level tasks. For diffusion model based techniques, we compare against DiffPath (6D) [10] and the Negative Log likelihood (NLL), via its Evidence Lower Bound (ELBO) when NLL cannot be acquired; these methods are only suitable for dataset-level OOD detection. We closely follow the official implementations of [10] and [35] to compute DiffPath (6D) and NLL respectively; full details of these implementations are provided in Section 8 and in our code. We also compare against the raw KL divergence in Equation (11) (denoted as D_{KL}), without our timestep restrictions; D_{KL} treats the entire image as one block for dataset-level OOD detection and considers each pixel as a separate block for image-level OOD detection. Additionally, we compare against two state of the art methods for anomaly detection that are not based on diffusion models: CutPaste [22] and SimpleNet [24]. We emphasize that CutPaste and SimpleNet operate directly on the true image, rather than taking indirect (CT or blurred) measurements as the diffusion-based methods do; this makes their OOD detection task somewhat easier.

We report OOD detection performance using the Area Under the receiver-operating characteristic Curve (AUC). For the dataset-level OOD detection task, we consider an image to be OOD if its KLIP value exceeds a threshold at any block of the image. For the image-level OOD detection task, we consider a block to be OOD if its KLIP value exceeds a threshold and report AUC evaluated over all blocks across all images in the *OOD set*.

Results. Our main results are reported in Table 1; computation time is reported in Section 11. KLIP applied to the predictor-corrector based diffusion model [37] outperforms all baselines for both dataset and image-level OOD detection, across both the star artifacts in the *Tuning set* and the realistic synthetic liver tumors in the *OOD set*, in the context of sparse-view CT reconstruction. This localization capability extends to the patch-based diffusion model [14]

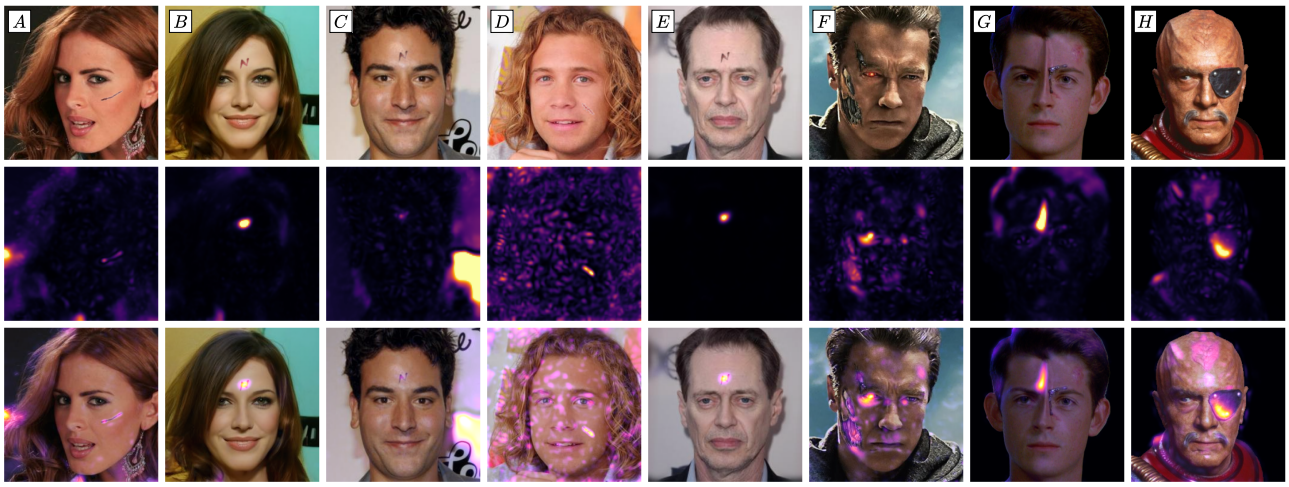

Figure 5. **Visual results for image-level OOD detection on human faces.** *Top:* Samples of the *OOD set* containing different localized artifacts such as scars. *Middle:* Heatmaps of KLIP computed pixel-wise (i.e. $D_B = 1$). *Bottom:* Heatmaps overlaid on images, to show localization. Columns A through E are images from the CelebA test set with synthetically-added scars; columns F through H are images from film and television of characters with distinctive and localized facial features.

using the same hyperparameters, and to detection of both scars and character features in the context of human face de-blurring with a DDPM-based model [41].

Figure 4 visualizes examples of OOD feature localization using KLIP, for CT scans in the *Tuning set* (with stars) and *OOD set* (with tumors) and for both the predictor-corrector based model (first 4 columns) and the patch-based model (last 4 columns). KLIP heatmaps largely coincide with the synthetic stars and tumors in the CT scans. While some blocks that do not contain OOD features also have high KLIP values, these values are consistently high for blocks that do contain local OOD features.

Similarly, Figure 5 visualizes examples of OOD feature localization using KLIP with a DDPM-based model [41] trained on CelebA [23], for Gaussian de-blurring of human faces with localized unusual (OOD) features. The first five columns show images from the CelebA test set, with added scars (inspired by Harry Potter's); for these examples we can ensure that only the added scar is OOD. The rightmost three columns show examples of film and television characters with unusual facial features such as facial prosthetics. No postprocessing is done on these images, but they may contain subtle image-wide distribution shifts compared to CelebA, in addition to the localized facial prosthetics. Across both types of face OOD detection, we find that KLIP successfully localizes the OOD features, though in some cases it also highlights other parts of the image.

Ablation Study. We perform three ablation experiments. First, we show that both block and timestep restrictions of Equation (11) are necessary to achieve both *dataset-level* and *image-level* detection of local OOD features. As in our main experiments, we select hyperparameters that maximize performance over the *Tuning set* (with stars) and use the same values on the *OOD set* (with tumors). Table 2 summarizes the resulting AUCs. We observe that restrictions to blocks or timestep ranges individually tend to improve performance, and that both restrictions (i.e., KLIP) usually perform better than either restriction alone.

Second, we present a sensitivity analysis of AUC perfor-

Table 2. **Ablation study over separate block and timestep restriction.** All AUC values are reported for the predictor-corrector based model. Hyperparameters are optimized over the *Tuning set* (star); the corresponding tuned AUC values are marked with †. The best AUC for each task is underlined.

	Dataset-Level		***Image-Level***	
	Star	Tumor	Star	Tumor
D_{KL}	0.54†	0.60	0.84†	0.86
+ block	0.85†	0.65	0.88†	<u>0.91</u>
+ time	0.57†	<u>0.78</u>	0.86†	0.81
KLIP	<u>0.86†</u>	<u>0.78</u>	<u>0.91†</u>	0.88

Table 3. **Sensitivity to hyperparameters.** Our main results are obtained by tuning hyperparameters on the *Tuning set* (stars) for the predictor-corrector based diffusion model. Here we compare AUC values when hyperparameters are tuned instead on the *OOD set* (tumors), to evaluate hyperparameter sensitivity. Tuned values are marked with †. The best AUC for each task is underlined.

	Dataset-Level		***Image-Level***	
	Star	Tumor	Star	Tumor
Tune on star	<u>0.86†</u>	0.78	<u>0.91†</u>	0.88
Tune on tumor	0.68	<u>0.87†</u>	0.85	<u>0.92†</u>

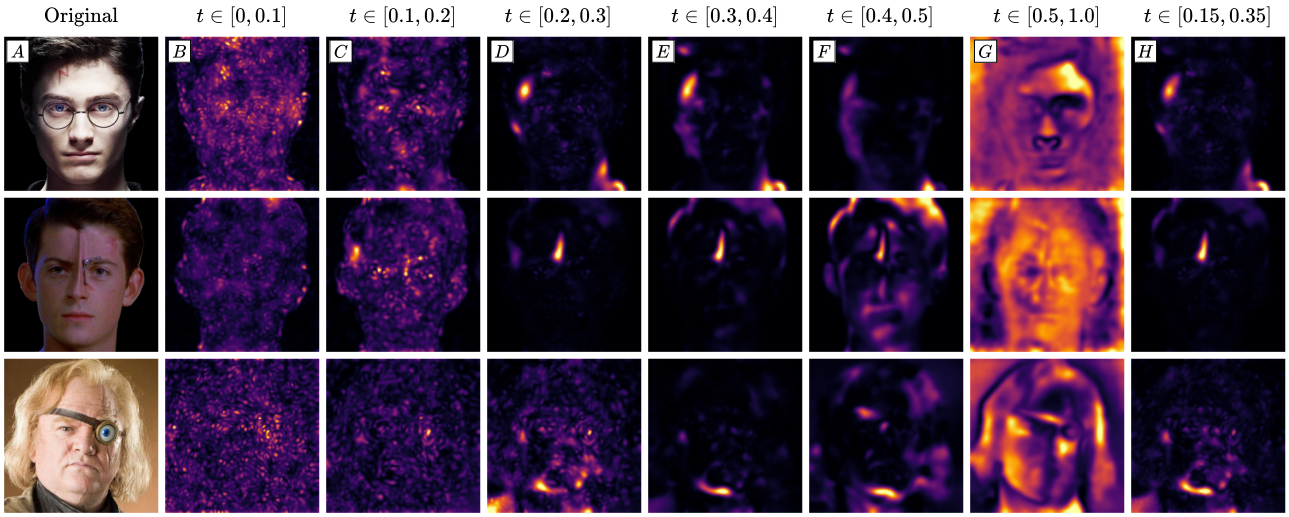

Figure 6. **Effect of different timestep ranges on KLIP.** (A) Original images before Gaussian blur is applied. (B-F) KLIP heatmaps for timestep windows of length 0.1, $[t_0, t_0 + 0.1]$, where t_0 starts at 0 in (B) and is incrementally increased in steps of 0.1 up to 0.4. (G) KLIP heatmap for the window $t \in [0.5, 1.0]$. (H) KLIP heatmap for the window $t \in [0.15, 0.35]$, which was tuned to optimize detection of star artifacts in sparse-view CT using a different diffusion model. We observe that different features are detected at different timesteps, in line with expectations that diffusion models add higher-frequency details later in sampling (at smaller values of t) [30, 48].

mance with respect to the choice of block size and timestep range. Table 3 shows how the AUCs change when the hyperparameters are optimized to maximize performance over the *OOD set* instead of the *Tuning set*. We find that tuning performance on one OOD dataset can reduce performance on the other; future investigation into robust hyperparameter selection should improve performance even further.

Third, in Figure 6, we qualitatively visualize KLIP for three OOD face images across varying timestep ranges and observe that different time ranges capture OOD features of varying sizes and shapes. At the earliest timesteps (large ts), KLIP highlights low-frequency components of the image such as the overall shape of the face. Smaller, more localized OOD features become more prominent as sampling proceeds (small ts). For example, Harry Potter's scar (top row), Icheb's metal implant (middle row), and Mad-Eye Moody's eye (bottom row) are all visible in panel C, while Moody's eye and Icheb's ridged nose, which are both larger, first appear earlier in panels D and F, respectively. This aligns with the observations of previous works [30, 48] that diffusion models construct low-frequency signals in the early stage of sampling, and progressively add fine details throughout sampling. It also explains why the hyperparameters $[t_0, t_1]$ that best detect local distribution shifts depend on the characteristics and scale of the OOD features.

6. Discussion

We focus on the understudied and challenging task of detecting spatially localized OOD features in images, given access only to indirect measurements of those images in the context of an inverse problem. We propose an OOD detection metric, KLIP, based on the spatially and temporally restricted KL divergence between the prior and posterior distributions, which we can estimate efficiently during posterior sampling with a diffusion model. This metric outperforms existing diffusion-based OOD detection methods across diverse imaging inverse problems (sparse-view CT and Gaussian deblurring) and inlier distributions (abdominal CT scans and human faces), for both dataset-level and image-level OOD detection of spatially localized OOD features. It also transfers across diffusion models, including both whole-image and patch-based architectures.

Limitations and Future Work. While our results are promising, they are also subject to the following limitations. First, KLIP is sensitive to hyperparameters for block size and time window, and could likely be improved with a more refined tuning strategy. We posit that this sensitivity is due to variation in the size and shape of real localized OOD artifacts, which influences the optimal block sizes and time windows for detection. Second, our evaluation framework is technically an *inverse crime*, because we use the same forward models to simulate the measurements (CT projections and blurred images) and perform posterior sampling reconstruction. While this provides a fair testbed to compare different OOD detection methods, and we show some robustness to model misspecification in Section 10, we hope that future work will validate and refine KLIP on real-world data for diverse local OOD detection tasks.

Acknowledgment

This work was supported in part by the HIVES program at Georgia Tech Research Institute (GTRI).

References

[1] Jonas Adler. Deep bayesian inversion computational uncertainty quantification for large scale inverse problems. 2018. 2

[2] Jonas Adler and Ozan Öktem. Deep posterior sampling: Uncertainty quantification for large scale inverse problems. 2019. 2

[3] Anastasios N Angelopoulos, Amit P Kohli, Stephen Bates, Michael I Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging, 2022. 1, 2

[4] Anastasios Nikolas Angelopoulos, Stuart R Pomerantz, Synho Do, Stephen Bates, Christopher P Bridge, Daniel C Elton, Michael H Lev, R Gilberto Gonzalez, Michael I Jordan, and Jitendra Malik. Conformal triage for medical imaging ai deployment. *medRxiv*, pages 2024–02, 2024. 1, 2

[5] Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection, 2024. 2

[6] Christopher M Bishop. Novelty detection and neural network validation. In *International Conference on Artificial Neural Networks*, pages 789–794. Springer, 1993. 2

[7] Stanley Chan et al. Tutorial on diffusion models for imaging and vision. *Foundations and Trends® in Computer Graphics and Vision*, 16(4):322–471, 2024. 3

[8] Siyi Chen, Huijie Zhang, Minzhe Guo, Yifu Lu, Peng Wang, and Qing Qu. Exploring low-dimensional subspace in diffusion models for controllable image editing. *Advances in neural information processing systems*, 37:27340–27371, 2024. 2, 5

[9] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 3, 1

[10] Alvin Heng, Harold Soh, et al. Out-of-distribution detection with a single unconditional diffusion model. *Advances in Neural Information Processing Systems*, 37:43952–43974, 2024. 1, 2, 6

[11] Matthias Hermann, Georg Umlauf, Bastian Goldluecke, and Matthias O Franz. Fast and efficient image novelty detection based on mean-shifts. *Sensors*, 22(19):7674, 2022. 2

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[13] Eliahu Horwitz and Yedid Hoshen. Conffusion: Confidence intervals for diffusion models, 2022. 2

[14] Jason Hu, Bowen Song, Xiaojian Xu, Liyue Shen, and Jeffrey A Fessler. Learning image priors through patch-based diffusion models for solving inverse problems. *Advances in Neural Information Processing Systems*, 37:1625–1660, 2024. 2, 5, 6

[15] Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan L Yuille, and Zongwei Zhou. Label-free liver tumor segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7422–7432, 2023. 2, 5

[16] Huiqun Huang, Sihong He, and Fei Miao. Cuqds: Conformal uncertainty quantification under distribution shift for trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17422–17430, 2025. 2

[17] Xinchao Jiang, X. F. Wanga, Ziming Wena, Enying Li, and Hu Wang. An e-pinn assisted practical uncertainty quantification for inverse problems. *ArXiv*, abs/2209.10195, 2022. 2

[18] Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data, 2019. 5, 2

[19] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in neural information processing systems*, 35:23593–23606, 2022. 3

[20] Gilad Kutiel, Regev Cohen, Michael Elad, Daniel Freedman, and Ehud Rivlin. Conformal prediction masks: Visualizing uncertainty in medical imaging. In *International Workshop on Trustworthy Machine Learning for Healthcare*, pages 163–176. Springer, 2023. 1, 2

[21] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 2

[22] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674, 2021. 6, 1, 3, 5, 7, 8, 9

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 5, 7

[24] 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*, pages 20402–20411, 2023. 6, 1, 3, 5, 7, 8, 9

[25] Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling. In *International conference on machine learning*, pages 15650–15665. PMLR, 2022. 1, 2

[26] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *International Conference on Learning Representations*, 2019. 2

[27] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. *arXiv preprint arXiv:1906.02994*, 2019. 2

[28] Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003. 5

[29] Nicolai André Brogaard Riis, Amal Alghamdi, Felipe Uribe, Silja L Christensen, Babak Maboudi Afkham, Per Christian Hansen, and Jakob Sauer Jørgensen. Cuqipy: I. computational uncertainty quantification for inverse problems in python. *Inverse Problems*, 40, 2023. 2

[30] Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. *International Conference on Learning Representations*, 2023. 2, 5, 8

[31] Hannah M. Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization, 2022. 2

[32] Shirin Shoushtari, Edward P Chandler, Yuanhao Wang, M Salman Asif, and Ulugbek S Kamilov. Unsupervised detection of distribution shift in inverse problems using diffusion models. *arXiv preprint arXiv:2505.11482*, 2025. 1, 2

[33] Shirin Shoushtari, Yi Wang, Xiao Shi, M Salman Asif, and Ulugbek S Kamilov. Eigenscore: Ood detection using covariance in diffusion models. *arXiv preprint arXiv:2510.07206*, 2025. 1, 2

[34] Jorge Silva and Shrikanth S Narayanan. Information divergence estimation based on data-dependent partitions. *Journal of Statistical Planning and Inference*, 140(11):3180–3198, 2010. 5

[35] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3, 5, 6, 1

[36] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34: 1415–1428, 2021. 4

[37] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021. 2, 3, 4, 5, 6, 1

[38] Julian Tachella and Marcelo Pereyra. Equivariant bootstrapping for uncertainty quantification in imaging inverse problems. In *International Conference on Artificial Intelligence and Statistics*, 2023. 2

[39] Jacopo Teneggi, Matthew Tivnan, J. Webster Stayman, and Jeremias Sulam. How to trust your diffusion model: A convex optimization approach to conformal risk control, 2023. 2

[40] Anton Vasiliuk, Daria Frolova, Mikhail Belyaev, and Boris Shirokikh. Limitations of out-of-distribution detection in 3d medical image segmentation. *Journal of Imaging*, 9(9):191, 2023. 2

[41] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 2, 5, 6, 7

[42] Hengkang Wang, Xu Zhang, Taihui Li, Yuxiang Wan, Tiancong Chen, and Ju Sun. Dmplug: A plug-in method for solving inverse problems with diffusion models. *Advances in Neural Information Processing Systems*, 37:117881–117916, 2024. 3

[43] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9):3064–3074, 2005. 5

[44] Jeffrey Wen, Rizwan Ahmad, and Philip Schniter. Task-driven uncertainty quantification in inverse problems via conformal prediction. *ArXiv*, abs/2405.18527, 2024. 2

[45] Tailin Wu, Willie Neiswanger, Hongtao Zheng, Stefano Ermon, and Jure Leskovec. Uncertainty quantification for forward and inverse problems of pdes via latent global evolution. *ArXiv*, abs/2402.08383, 2024. 2

[46] Zihui Wu, Yu Sun, Yifan Chen, Bingliang Zhang, Yisong Yue, and Katherine Bouman. Principled probabilistic imaging using diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 37: 118389–118427, 2024. 3

[47] Xingyu Xu and Yuejie Chi. Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction. *Advances in Neural Information Processing Systems*, 37:36148–36184, 2024. 3

[48] Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 22552–22562, 2023. 5, 8

[49] Ximiao Zhang, Min Xu, Dehui Qiu, Ruixin Yan, Ning Lang, and Xiuzhuang Zhou. Mediclip: Adapting clip for few-shot medical image anomaly detection, 2024. 2

[50] Yifei Zong, David A. Barajas-Solano, and Alexandre M. Tartakovsky. Randomized physics-informed machine learning for uncertainty quantification in high-dimensional inverse problems. *ArXiv*, abs/2312.06177, 2023. 2