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

Supplementary Material

7. Likelihood Score Approximation

We evaluate KLIP in the context of two prior works [9, 37] with different posterior sampling algorithms. In [37], an additional step at each time t replaces the sample x_t with x'_t, which is the solution of a proximal optimization step to ensure consistency of the sample with the measurement y. Specifically, x'_t is the solution of the optimization problem

$$x'_t = \arg\min_{z\in\mathbb{R}^D}\{(1-\lambda)\|z-x_t\|_B^2 + \min_{u\in\mathbb{R}^D}\|z-u\|_B^2\} \quad s.t.\ \ Au = y_t, \tag{13}$$

where A is a linear forward model that involves an invertible square matrix B, y_t is the simulated measurement using the current sample x_t, and λ is a hyperparameter that controls how strongly the measurements should affect the sampling process. We refer interested readers to [37] for more details. While this work does not directly approximate the likelihood score, the update made by the proximal step is precisely the difference between unconditional sampling (Equation (3)) and posterior sampling (Equation (6)), which can be expressed as the following equation:

$$-g(t)^2 s_l(x_t, y; t) \simeq x'_t - x_t. \tag{14}$$

Therefore, we consider the scaled update

$$s_l(x_t, y; t) \simeq \frac{x_t - x'_t}{g(t)^2} \tag{15}$$

as an implicit approximation of the likelihood score, and use it to compute KLIP.

On the other hand, [9] approximates the score with

$$\nabla_{x_t}\log p(y|x_t) \simeq -\frac{1}{\sigma^2}\nabla_{x_t}\|y - \mathcal{A}(\hat{x}_0(x_t))\|_2^2, \tag{16}$$

where $\mathcal{A}$ is the forward model, $\hat{x}_0(x_t) = \mathbb{E}[x_0|x_t]$ is the predicted state at $t = 0$ given the current state x_t, σ^2 is the variance of the Gaussian noise included in the forward model, and ζ_t is a parameter that controls how strong the likelihood affects the sampling process.

8. Baseline Computation

We compare KLIP against 2 primary baselines, NLL (negative log likelihood) and DiffPath [10]. For NLL, we directly use the official implementation on github by the authors of [35]. It computes the exact likelihood instead of the Evidence Lower Bound (ELBO) using the probability flow ODE, which is an ordinary differential equation whose marginals coincide with the underlying SDE of the diffusion model.

DiffPath [10] constructs a 6-dimensional feature vector for each image, and uses these features to determine if the image is OOD. These features are the ℓ_1, ℓ_2, and ℓ_3 norms of $\epsilon_\theta(x_t, t)$, defined as:

$$\epsilon_\theta(x_t, t) = -\sigma(t)s_\theta(x_t, t), \tag{17}$$

and its time derivative. Here, $\sigma(t)$ is a function used to define the diffusion coefficients $g(t)$ of the SDE. Specifically, both diffusion models that we evaluate KLIP on are defined by a class of SDEs called Variance Exploding (VE) SDEs, which take the form

$$dx = \sqrt{\frac{d\sigma^2(t)}{dt}}dw. \tag{18}$$

To evaluate $\epsilon_\theta(x_t, t)$, the score model is queried once every timestep. Since the predictor-corrector based model [37] involves multiple updates by the predictor and the corrector at each timestep, we query the score model once after the last corrector update for each timestep. Additionally, DiffPath fits a Kernel Density Estimator (KDE) to the feature vectors of the images in the training set. For both models, we instead fit DiffPath's KDE to feature vectors extracted from images in the ID validation set, that do not overlap with images in the ID training or evaluation sets. We used 250 images to fit KDE for the predictor-corrector model, and 100 images for the patch-based model.

9. Robustness Evaluation

We perform an additional comparison to evaluate the generalizability of KLIP to local OOD features with different properties, specifically simulated liver tumors with different sizes and densities. Figure 7 shows sample images from the different OOD sets we used for this evaluation, acquired by modifying the darkness and size of the simulated liver tumors. The OOD set that we used for evaluation in the main manuscript corresponds to the 6th row, containing large tumors with medium darkness.

We also compare KLIP against two additional baselines: CutPaste [22] and SimpleNet [24], using their public code and default parameter settings (with the exception of image size, which we adjust to match our dataset). While these works mainly aim to detect anomalies or defects in the context of industrial visual inspection, they have been extended

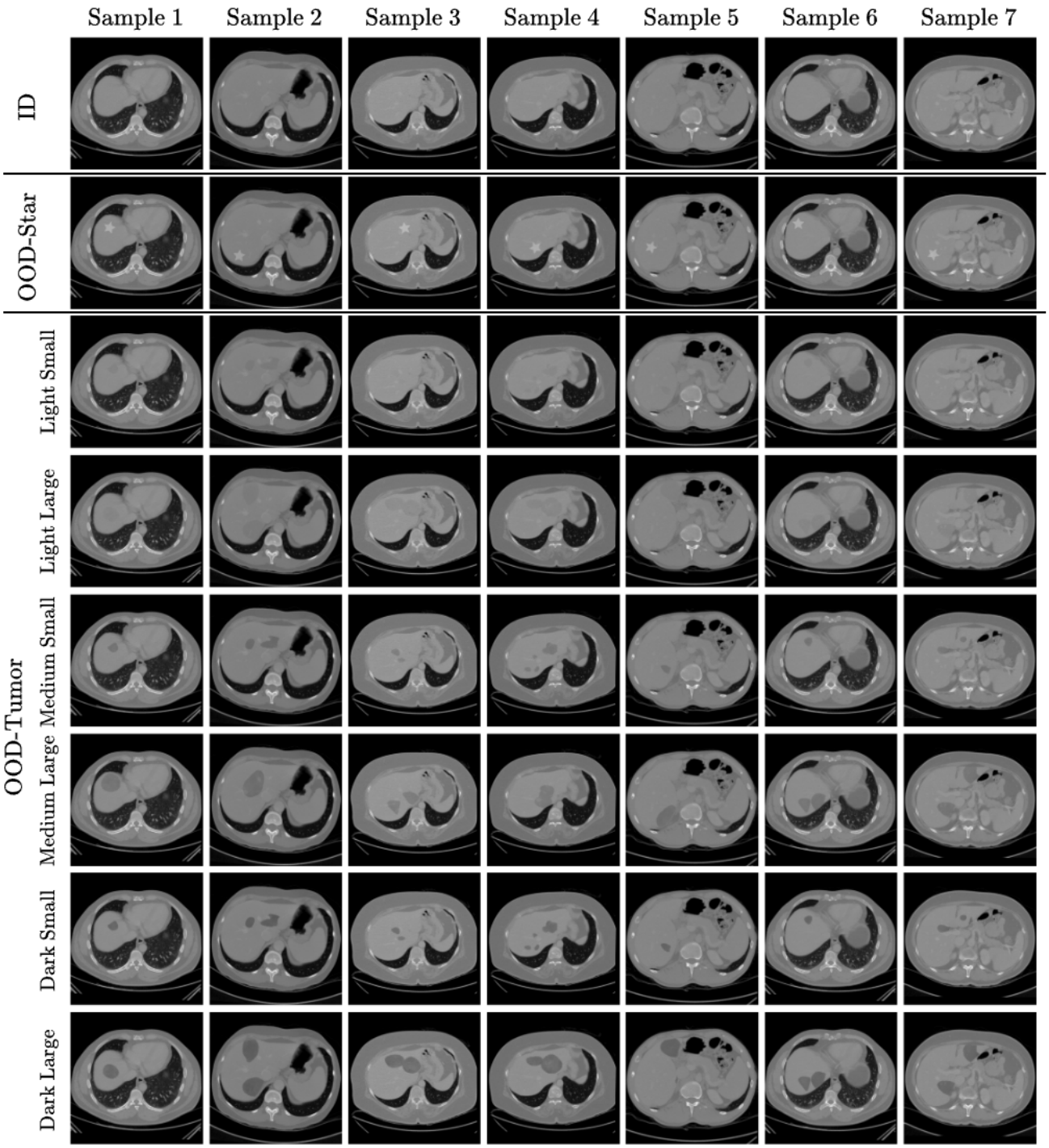

Figure 7. **Sample images included in the ID and different OOD sets.** Each row represents a distinct dataset, and each column shows 7 different samples from that dataset. From the top, we have (1) the ID set consisted of healthy CT scans from the CHAOS [18] evaluation dataset, (2) the OOD set with a star shaped artifact, used for hyperparameter tuning, and (3) 6 OOD sets containing synthetic liver tumors of different darkness and shape, generated following [15].

to other settings including medical imaging [5, 31, 49]. We train both CutPaste and SimpleNet with the training set of the CHAOS dataset [18], which we used to train both of our diffusion models.

The results are presented in Table 7. We find that dataset-level OOD detection performance is fairly stable across tumor types, for all models. KLIP yields the highest dataset-level AUC across all tumor types, when using the predictor-corrector based diffusion model. However, all dataset-level OOD metrics struggle when applied to the patch-based PaDIS diffusion model. Empirically, we observe that KLIP can achieve higher AUC metrics for PaDIS when its hyperparameters are chosen specifically for that model, so we are optimistic that future work may introduce a more adaptive hyperparameter selection strategy that would render KLIP more robust across different diffusion model architectures.

For image-level OOD detection, KLIP shows strong performance when applied to both the predictor-corrector based diffusion model and the patch-based diffusion model. In both cases, AUC metrics degrade gradually with decreasing tumor size and darkness. Figure 9 to Figure 10 visually compare the localization of baseline metrics and KLIP. Figure 9 shows examples of images with star artifacts, while the figures from Figure 10 to Figure 15 differ only in the size and darkness of the simulated tumors. All figures follow the same structure: the first row contains the sampled images, the second row shows the CutPaste [22] heatmap overlays, the third row presents the SimpleNet [24] heatmap overlays, and the fourth row displays the results of our method KLIP using the predictor–corrector diffusion model [37]. While the baseline methods accurately detect the star artifacts, they struggle to locate tumors of any size or darkness, instead highlighting other regions of the anatomy that are actually in-distribution. Although KLIP does not have perfect tumor detection either, it shows much stronger tumor localization than either baseline, and is especially adept at detecting darker tumors.

10. Forward Model Mismatch

Following prior work (e.g., [9, 37]), our main experiments assume a matched forward model, meaning that the forward model in the measurement and reconstruction are same. This setting allows us to evaluate KLIP without the additional effect of forward model misspecification. However, exact knowledge of the forward model may be unavailable in practice, and reconstruction often relies on only an approximate operator. To study the sensitivity of KLIP to model mismatch, we evaluate KLIP in the Gaussian deblurring task by varying the reconstruction blur kernel while keeping the true measurement blur kernel fixed. As shown in Figure 8 and Table 4, KLIP remains effective under such mismatch, although its precision gradually degrades as the assumed blur kernel moves away from the true one.

Table 4. Average image-level AUC of KLIP under reconstruction-model mismatch on 100 CelebA images with synthetic scar artifacts. The score is computed using 8 samples. The true measurement blur kernel is fixed at $\sigma = 9$, while the reconstruction blur kernel is varied. Performance is best under the matched setting and degrades gracefully under moderate mismatch.

σ_{rec}	3	7	9	11	13
AUC_{avg}	0.809	0.897	0.904	0.898	0.893

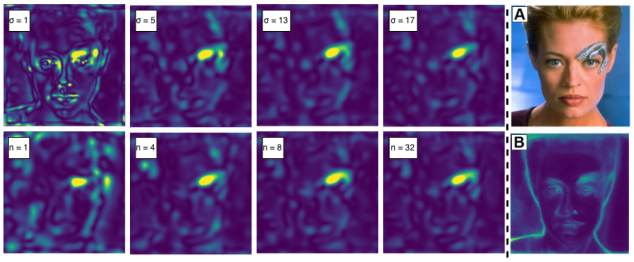

Figure 8. Left of the dotted line: KLIP heatmaps. Top: forward model mismatch—measurements use a 21×21 Gaussian blur kernel with $\sigma = 9$, while reconstruction uses kernels with $\sigma \in \{1, 5, 13, 17\}$ (KLIP computed with 8 samples). Bottom: effect of the number of samples $\{1, 4, 8, 32\}$ with no model mismatch. Right of the dotted line: (A) original image; (B) pixel-wise 95% conformal CI length for the sample mean over 128 samples.

Table 5. **Runtime comparison** in seconds. "Sampling" denotes the time required to generate 8 reconstructions at 512×512 from a single CT measurement y. The remaining entries report the total runtime of each method, including reconstruction and scoring.

Sampling	NLL	DiffPath	SimpleNet	Cutpaste	KLIP
375	518	548	376	376	383

11. Computational Evaluation

We also compare the computational cost of KLIP against baseline OOD detection methods in Table 5. We report the time to sample 8 reconstructions at 512×512 from a single CT measurement y using [37], and the total runtime of each method. KLIP takes only $\sim 2\%$ longer than sampling alone. Results indicates that KLIP achieves competitive detection performance while remaining computationally close to the underlying reconstruction pipeline.

12. Sample Size Sensitivity

Since KLIP estimates an expectation through Monte Carlo sampling, we evaluate how sensitive its performance is to the number of samples used in approximating the expectation. As shown in Figure 8 and Table 6, KLIP is reasonably stable across different sampling budgets, with performance improving as more samples are used, but with diminishing returns beyond a moderate number of samples. In particular, the average image-level AUC over 100 CelebA images

Table 6. Average image-level AUC of KLIP on 100 CelebA images with synthetic artifacts for different Monte Carlo sample sizes used to approximate the expectation in KLIP. Here, N denotes the number of samples, and AUC denotes the average image-level OOD detection performance. Performance improves with larger N, but the gain becomes marginal beyond 8 samples.

N	1	4	8	16	32
AUC	0.813	0.854	0.904	0.905	0.909

with synthetic scar artifacts increases from 0.813 with a single sample to 0.904 with 8 samples, and then changes only marginally for larger sample sizes. These results suggest that a moderate sampling budget already provides a reliable approximation, offering a favorable trade-off between computational cost and OOD detection performance.

Table 7. **AUC results for dataset-level and image-level OOD detection, for different OOD artifacts and models on a sparse-view CT inverse problem.** 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.

		Tumor						**Star**
		Light		**Medium**		**Dark**		
		Small	Large	Small	Large	Small	Large	
Dataset Level	CutPaste [22]	0.491	0.485	0.487	0.505	0.484	0.506	<u>0.999</u>
	SimpleNet [24]	0.499	0.501	0.505	0.504	0.499	0.501	0.993
	Predictor-Corrector [37]							
	NLL	0.511	0.514	0.528	0.535	0.540	0.559	0.586
	DiffPath	0.334	0.342	0.349	0.368	0.370	0.409	0.688
	D_{KL}	0.531	0.537	0.580	0.602	0.600	0.621	0.541
	KLIP (Ours)	<u>0.754</u>	<u>0.774</u>	<u>0.772</u>	<u>0.776</u>	<u>0.772</u>	<u>0.782</u>	0.855†
	PaDIS [14]							
	NLL	0.498	0.478	0.525	0.490	0.469	0.460	0.502
	DiffPath	0.469	0.482	0.507	0.497	0.504	0.481	0.480
	D_{KL}	0.498	0.534	0.510	0.545	0.538	0.643	0.506
	KLIP (Ours)	0.499	0.513	0.510	0.502	0.523	0.512	0.512
Image Level	CutPaste [22]	0.754	0.695	0.651	0.441	0.651	0.319	0.978
	SimpleNet [24]	0.378	0.216	0.822	0.592	0.920	0.657	<u>0.999</u>
	Predictor-Corrector [37]							
	D_{KL}	<u>0.799</u>	<u>0.800</u>	<u>0.890</u>	0.856	0.906	0.853	0.837
	KLIP (Ours)	0.785	0.791	0.876	<u>0.878</u>	0.920	0.904	0.912†
	PaDIS [14]							
	D_{KL}	0.595	0.663	0.857	0.672	<u>0.932</u>	<u>0.941</u>	0.841
	KLIP (Ours)	0.667	0.630	0.859	0.732	0.911	0.852	0.889

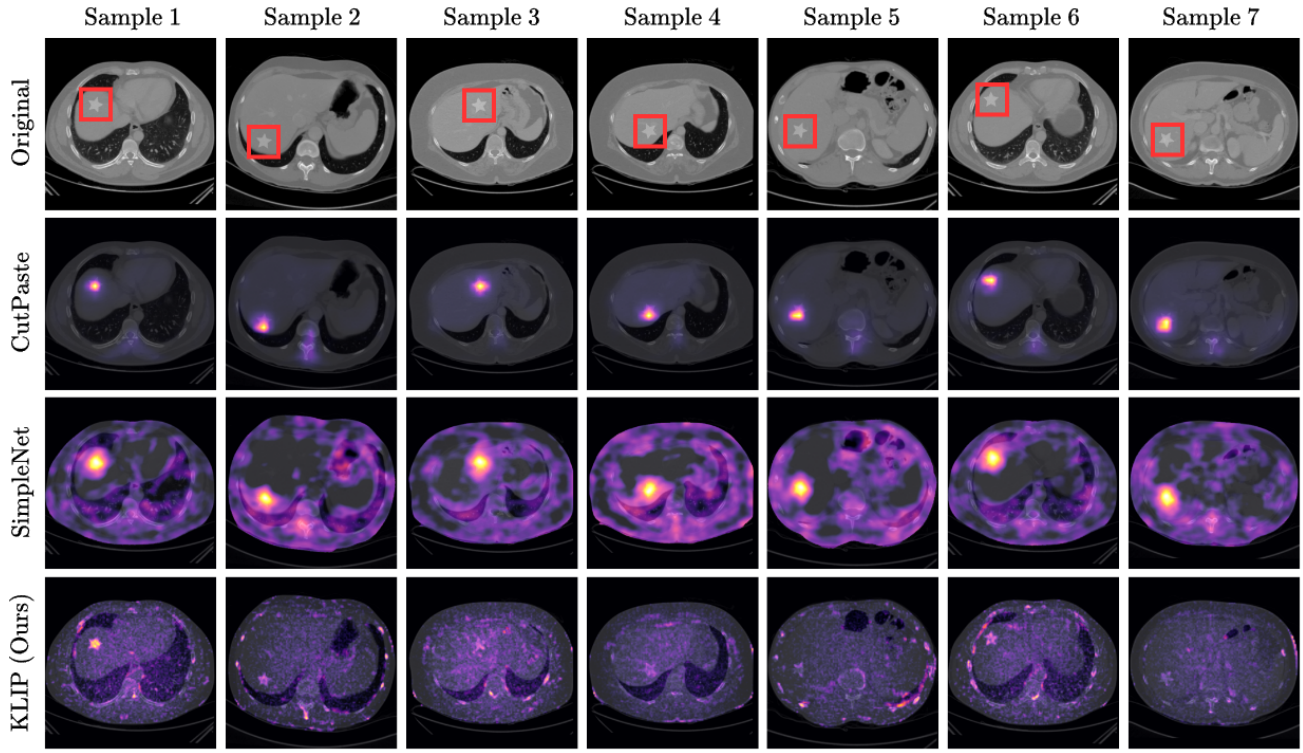

Figure 9. **Visual results for image-level OOD detection on sparse-view CT scans.** *Row 1:* Images in the OOD set with synthetic star artifacts. Red boxes annotate where the stars are. *Rows 2-4:* Heatmaps of CutPaste [22], SimpleNet [24], and KLIP overlaid on images.

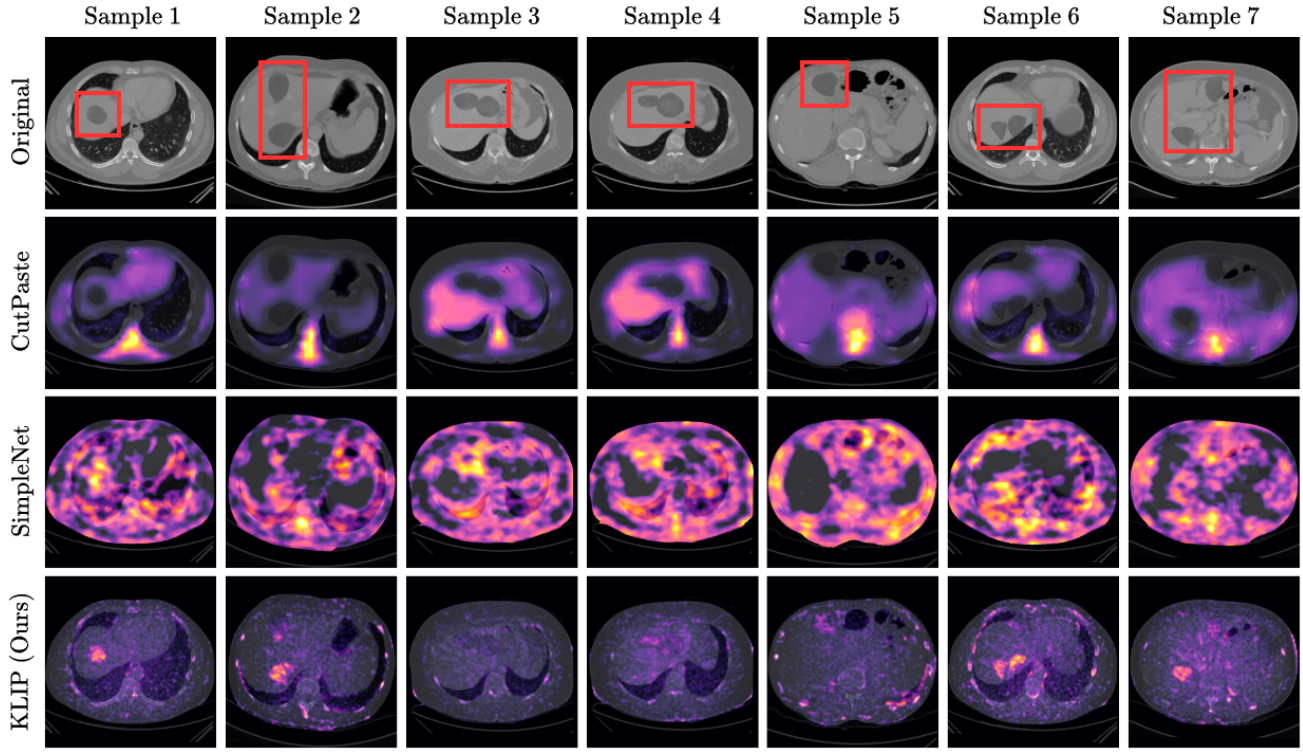

Figure 10. **Visual results for image-level OOD detection on sparse-view CT scans.** *Row 1:* Images in the OOD set with dark and large tumors. Red boxes annotate where the tumors are. *Rows 2-4:* Heatmaps of CutPaste [22], SimpleNet [24], and KLIP overlaid on images.

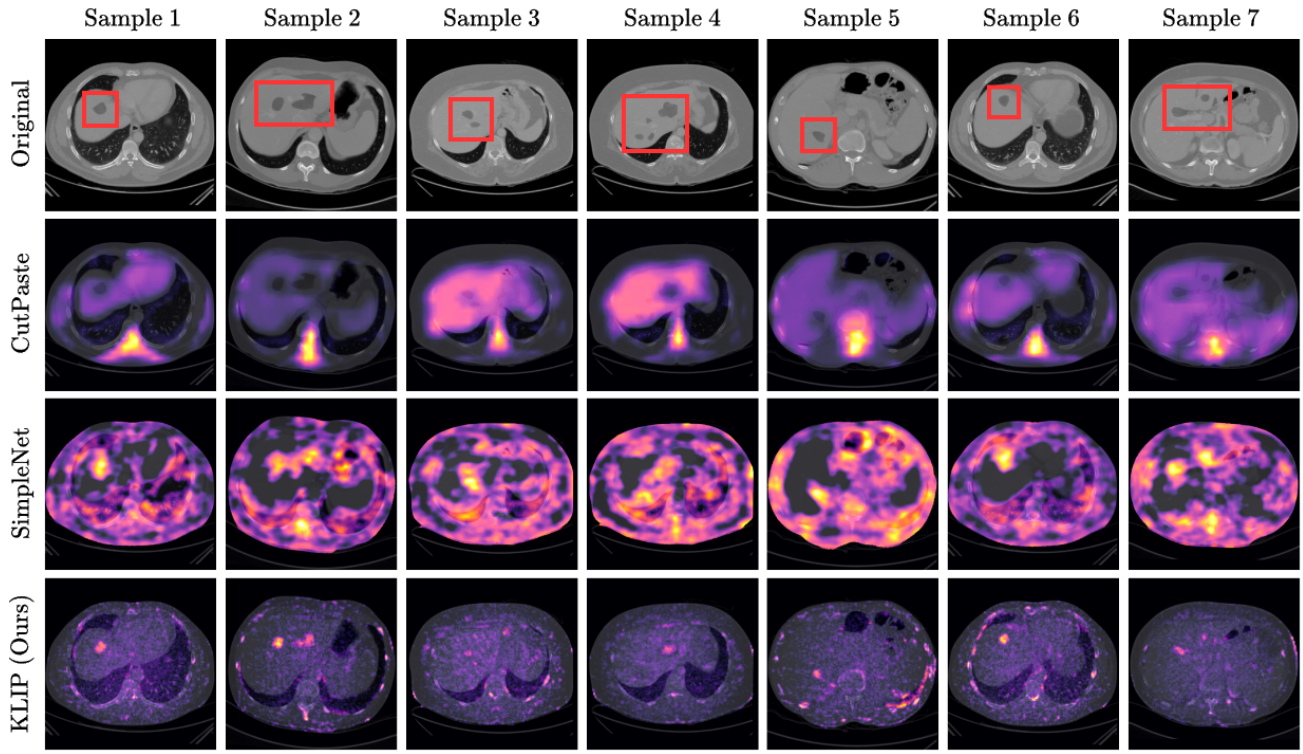

Figure 11. **Visual results for image-level OOD detection on sparse-view CT scans.** *Row 1:* Images in the OOD set with dark and small tumors. Red boxes annotate where the tumors are. *Rows 2-4:* Heatmaps of CutPaste [22], SimpleNet [24], and KLIP overlaid on images.

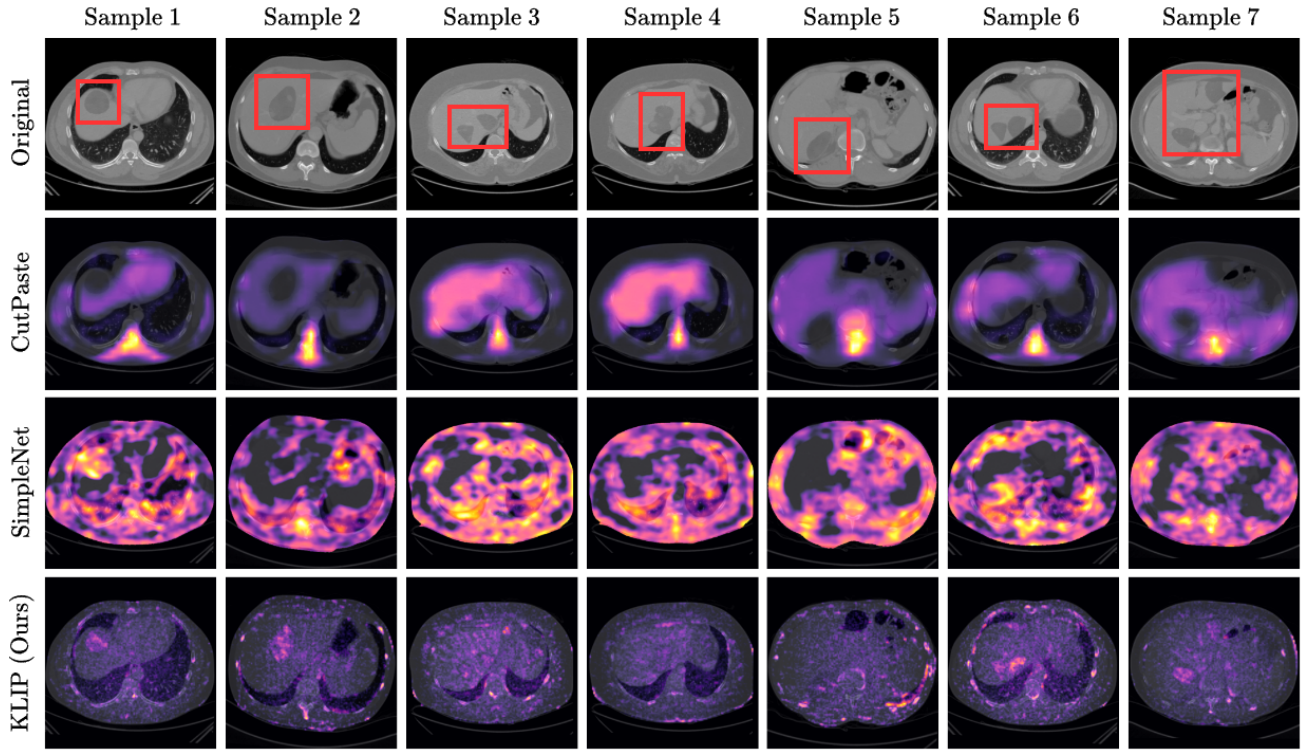

Figure 12. **Visual results for image-level OOD detection on sparse-view CT scans.** *Row 1:* Images in the OOD set with large tumors of medium darkness. Red boxes annotate where the tumors are. *Rows 2-4:* Heatmaps of CutPaste [22], SimpleNet [24], and KLIP overlaid on images.

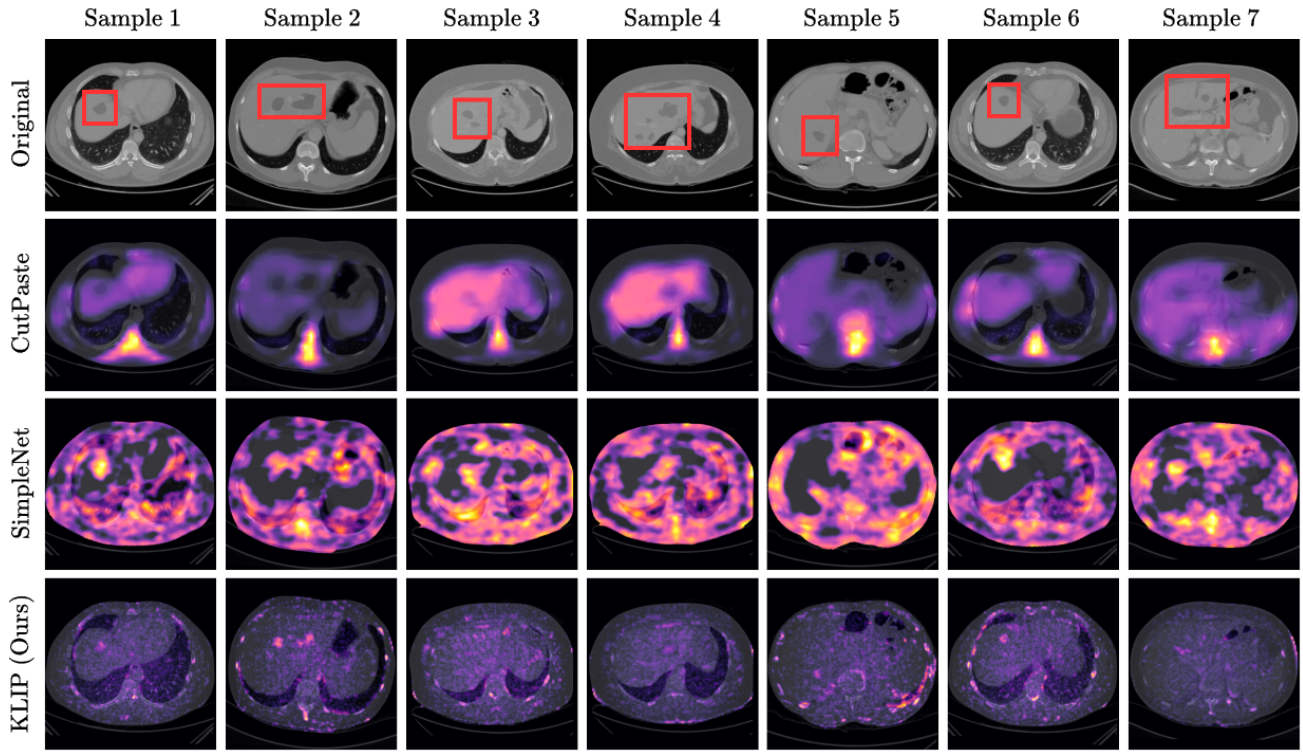

Figure 13. **Visual results for image-level OOD detection on sparse-view CT scans.** *Row 1:* Images in the OOD set with small tumors of medium darkness. Red boxes annotate where the tumors are. *Rows 2-4:* Heatmaps of CutPaste [22], SimpleNet [24], and KLIP overlaid on images.

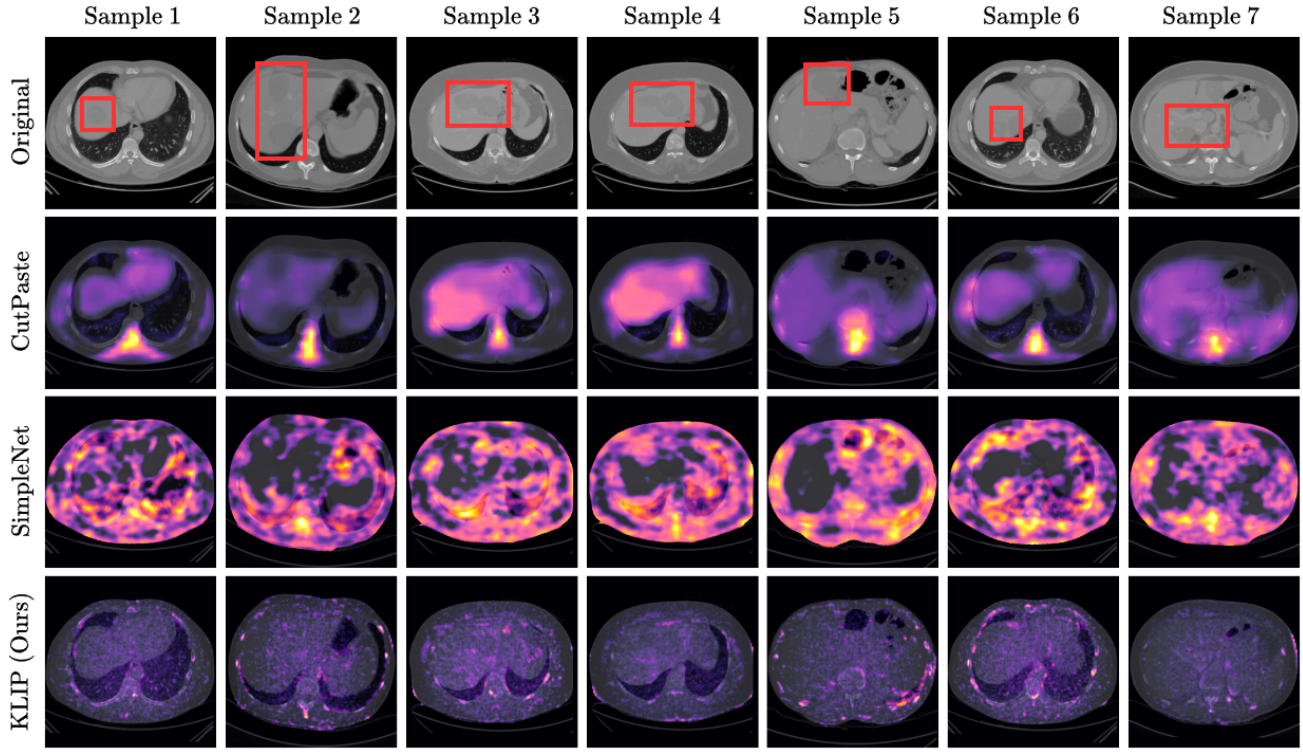

Figure 14. **Visual results for image-level OOD detection on sparse-view CT scans.** *Row 1:* Images in the OOD set with light and large tumors. Red boxes annotate where the tumors are. *Rows 2-4:* Heatmaps of CutPaste [22], SimpleNet [24], and KLIP overlaid on images.

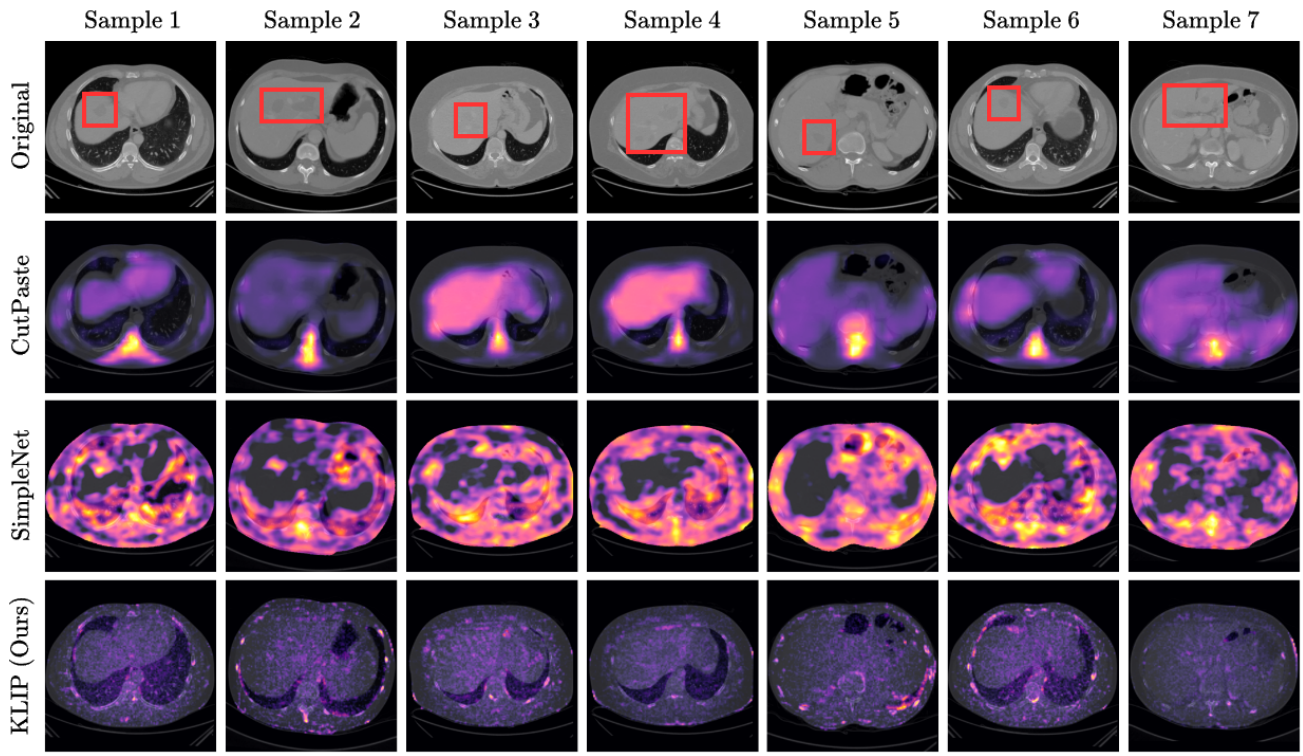

Figure 15. **Visual results for image-level OOD detection on sparse-view CT scans.** *Row 1:* Images in the OOD set with light and small tumors. Red boxes annotate where the tumors are. *Rows 2-4:* Heatmaps of CutPaste [22], SimpleNet [24], and KLIP overlaid on images.