Exploring Out-of-distribution Detection for Sparse-view Computed Tomography with Diffusion Models: Supplementary Experiments

Ezgi Demircan-Tureyen*, Felix Lucka, Tristan van Leeuwen

*edt@cwi.nl

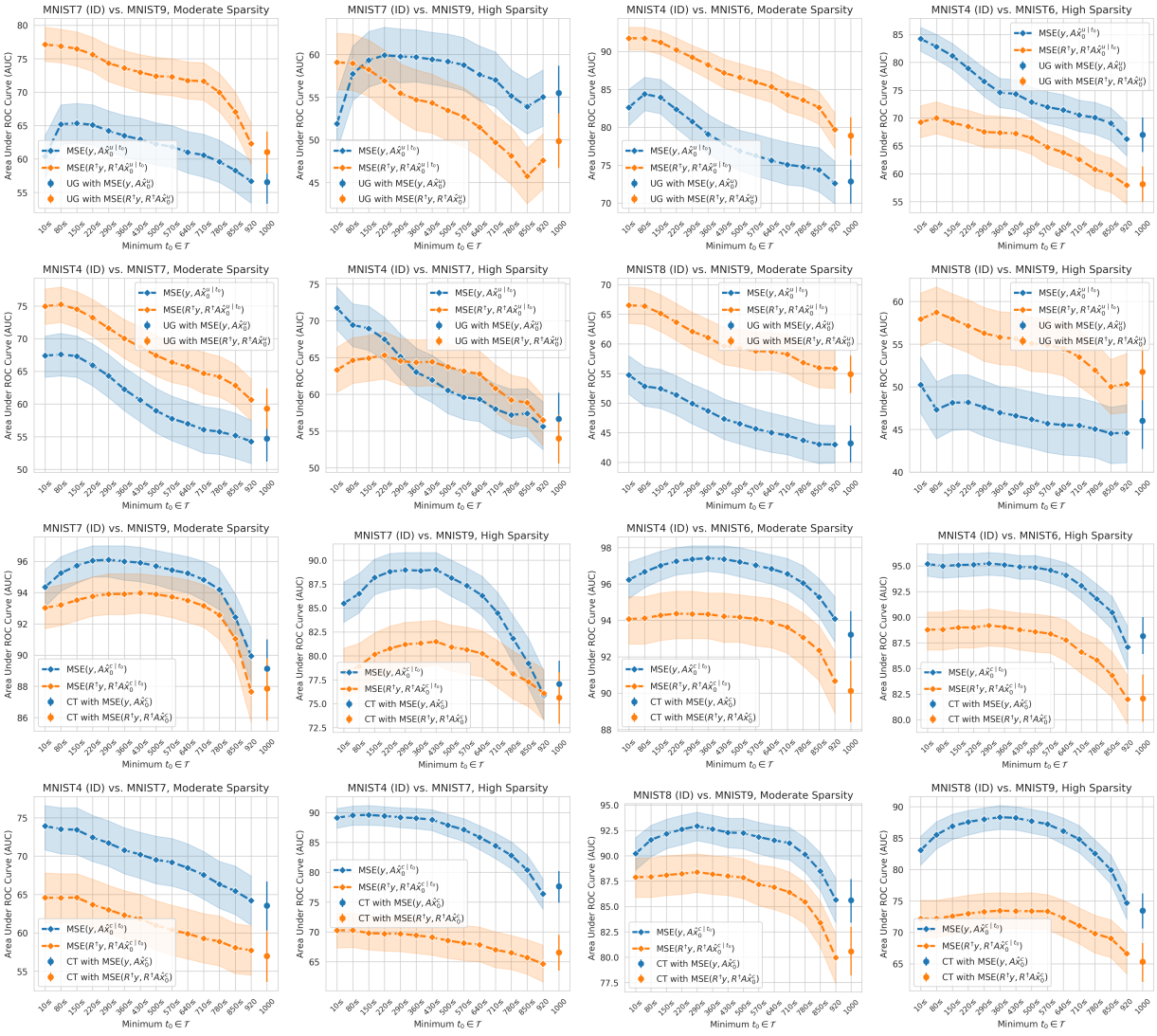


Figure S1: AUC scores with respect to different minimum values of t_0, shown for unconditional sampling **(top two rows)** and conditional sampling **(bottom two rows)**, across various "ID vs. OOD, Moderate/High Sparsity" scenarios, indicated in the individual plot titles, with $\{18, 12, 9\}$ and $\{6, 5, 4\}$ number of projections, respectively. As the minimum t_0 increases from 10 to 920, the number of reconstructions for OOD detection decreases from 14 to 1. The result shown at $t_0 = 1000$ represents what would be obtained without applying multi-scale scoring with FBP inputs. This corresponds to unconditional generation (denoted as "UG") and CT reconstruction (denoted as "CT") in unconditional **(top two rows)** and conditional cases **(bottom two rows)**, respectively. The shaded areas around the AUC values indicate confidence bounds, reported for all cases based on 1000-fold bootstrapping.

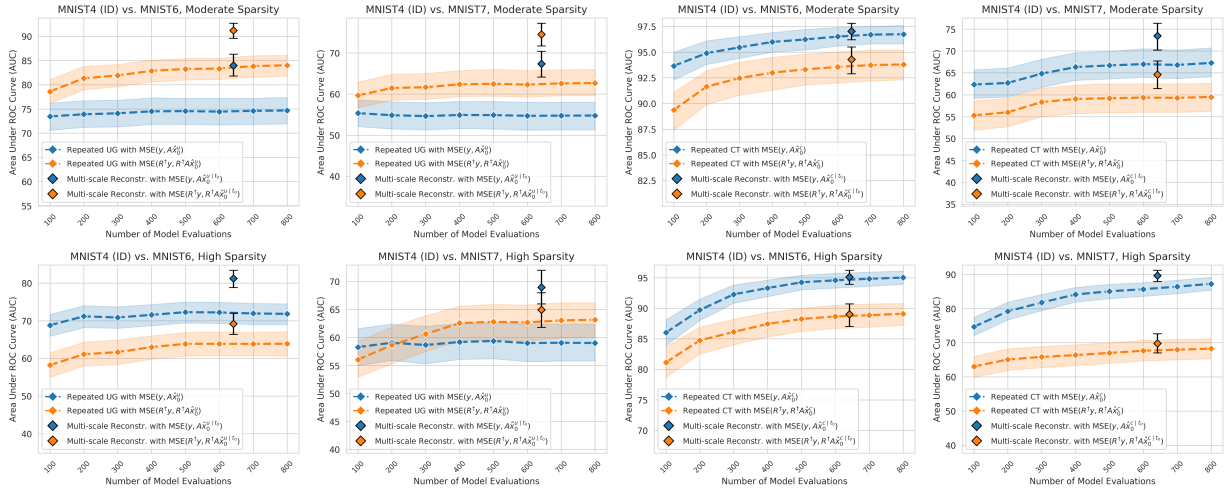
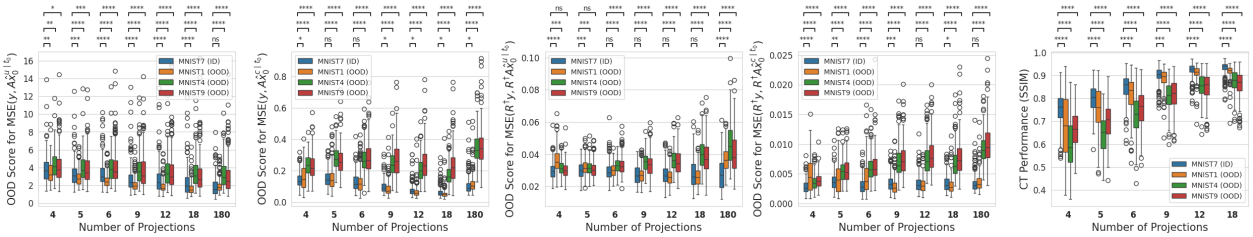
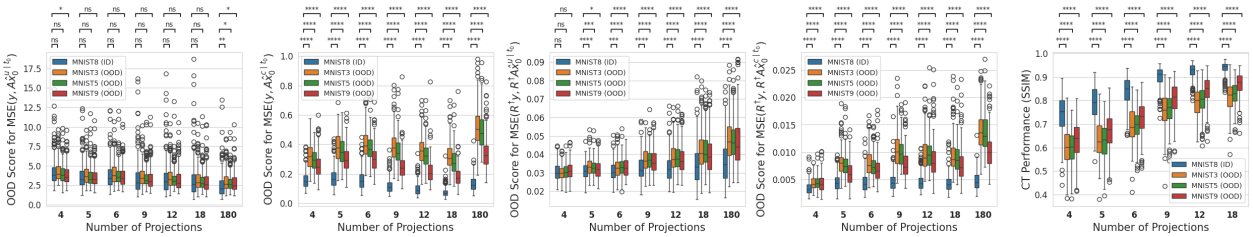


Figure S2: AUC scores as a function of the total number of model evaluations, where each evaluation count of $100 \times n$ represents running the generation process n times to mitigate randomness and ensure a fair comparison. The evolutions of both unconditional ("UG") and conditional ("CT", since it corresponds to CT reconstruction) generations are compared against the AUC scores obtained from multi-scale reconstructions by using noisy FBP inputs, depicted by individual error bars at 642 model evaluations. The analysis considers the model trained on MNIST4, with MNIST6 and MNIST7 treated as OOD to complement the visual results in Fig. 5. The **top row** presents results for Moderate Sparsity, while the **bottom row** corresponds to High Sparsity. Shaded regions around AUC values indicate confidence bounds, derived from 1000-fold bootstrapping. The use of 100 model evaluations per generation is due to the PLMS sampler.



(a) OOD score and SSIM distributions based on the model trained on MNIST7.



(b) OOD score and SSIM distributions based on the model trained on MNIST8.

Figure S3: This figure presents additional analyses complementary to Fig. 7. In both (a) and (b), the first two plots from the left represent Z-scores derived from MSE errors in the sinogram domain, with the first plot using unconditional reconstructions and the second using conditional reconstructions. Similarly, the third and fourth plots focus on the FBP domain, again using unconditional and conditional reconstructions, respectively. The final plot in (a) illustrates SSIM distributions to assess CT reconstruction ($t_0 = 1000$) performance rather than OOD detection. The statistical significance annotations indicated by asterisks denote p-values from t-tests for independent samples comparing ID and OOD distributions (* for $p \leq 5.00 \times 10^{-2}$, ** for $p \leq 1.00 \times 10^{-2}$, *** for $p \leq 1.00 \times 10^{-3}$, **** for $p \leq 1.00 \times 10^{-4}$, and ns for non-significant results).

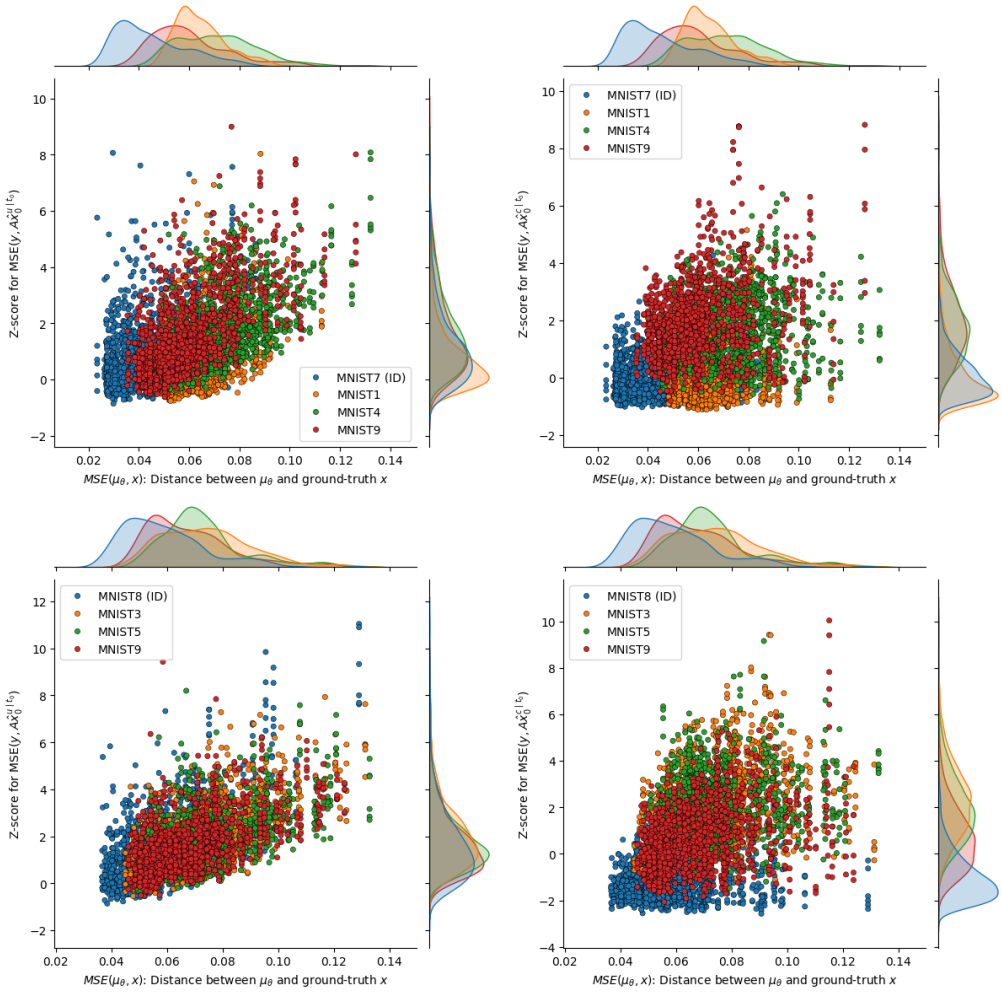


Figure S4: This figure presents additional analyses complementary to Fig. 8. Joint scatter plots illustrating how OOD scores vary with increasing distance of the image from the mean of the prior distribution, in unconditional (left column) and conditional (right column) cases, for the models trained on MNIST7 (top row) and MNIST8 (bottom row). For both cases, the positive correlation between the distance and the OOD score, initially moderate with r = 0.47 and r = 0.58 ($p < 0.0001$) when using MSE($y, A\hat{x}_0^{u|t_0}$), weakens to r = 0.31 and r = 0.41 ($p < 0.0001$) with MSE($y, A\hat{x}_0^{c|t_0}$). For the model trained on MNIST7, conditioning results in a substantial portion of MNIST1 images receiving lower OOD scores, similar to those of ID images. For the model trained on MNIST8, however, this effect is not observed. Conditioning compresses the ID scores to distinctly lower levels than those of all OOD sets, making them more distinguishable.

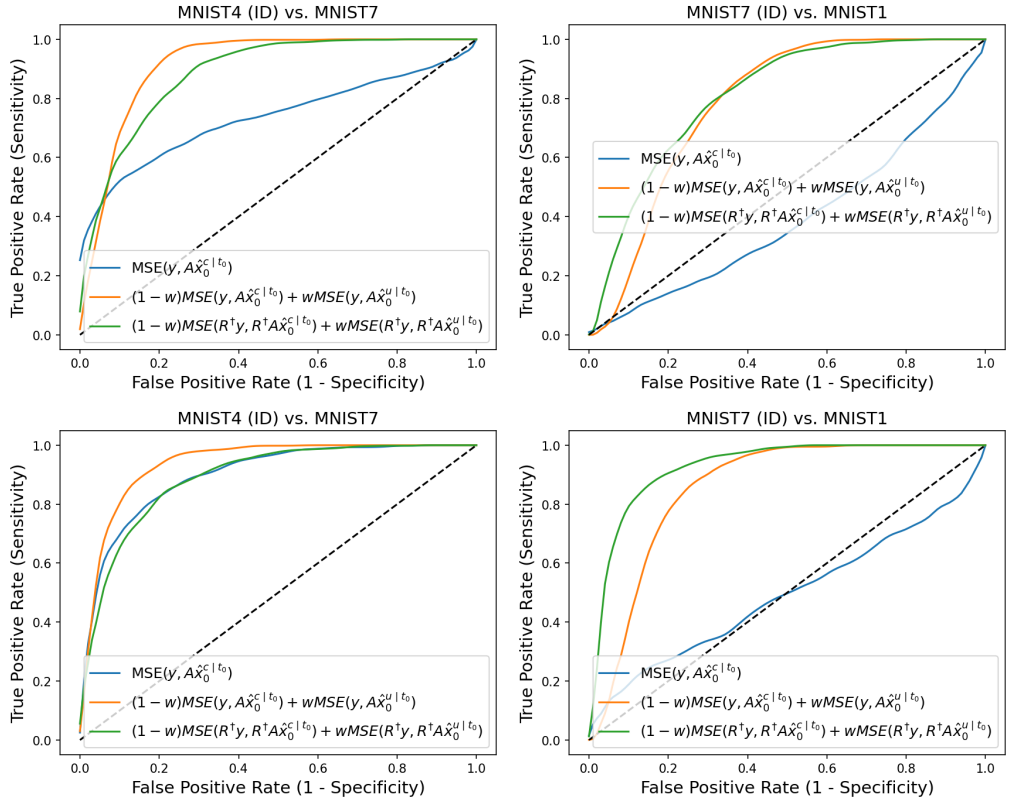


Figure S5: The ROC curves demonstrate the trade-off between TPR (sensitivity) and $1-\mathrm{FPR}$ (specificity) in cases where the proposed weighting scheme achieved a **higher** AUC than using reconstruction error in the projection domain with conditional samples. The top row plots the results for the Moderate Sparsity scenario (with $\{18, 12, 9\}$ projections), whereas the bottom row considers the High Sparsity scenario (with $\{6, 5, 4\}$ projections).

Table S1: AUC scores for three models trained on full-view MNIST4, MNIST7, and MNIST8 images, evaluated using two different distance metrics for the weighting scheme. Here, w_{μ_θ} represents the normalized L^2 distance between the measurement and the forward-projected mean of the training set, while w_{avg} corresponds to the normalized L^2 distance averaged across all training set projections. (i.e., $w = \frac{\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}}\|\mathbf{y}-\mathbf{A}x\|^2}{\|\mathbf{y}\|^2+\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}}\|\mathbf{A}x\|^2}$). Red texts highlight the highest values in each column separately for two cases: Moderate Sparsity, where the number of projection angles (#proj) is $\{18, 12, 9\}$, and High Sparsity, where #proj = $\{6, 5, 4\}$. Confidence bounds are reported for all cases using 1000-fold bootstrapping. This alternative distance metric is included as an illustrative example but is neither effective nor practical due to its high computational cost.

	ID Dataset:	MNIST4			MNIST7			MNIST8		
	OOD Dataset:	MNIST6	MNIST7	MNIST9	MNIST1	MNIST4	MNIST9	MNIST3	MNIST5	MNIST9
Moderate	w_{μ_θ} $(\Gamma = \mathbf{I})$	96.25 [95.2, 97.2]	**91.45** [89.8, 93.1]	76.51 [73.6, 79.3]	78.33 [75.6, 80.9]	88.06 [86.0, 89.9]	83.95 [81.5, 86.1]	87.79 [85.6, 89.9]	87.72 [85.5, 89.9]	81.78 [79.2, 84.4]
	w_{μ_θ} $(\Gamma = \mathbf{R}^\dagger)$	**97.98** [97.3, 98.6]	88.46 [86.6, 90.3]	**83.60** [81.4, 85.9]	**81.07** [78.8, 83.3]	**94.20** [92.9, 95.5]	**91.39** [89.8, 93.0]	**94.58** [93.3, 95.7]	**95.03** [93.8, 96.2]	**90.34** [88.6, 91.9]
	w_{avg} $(\Gamma = \mathbf{I})$	95.52 [94.4, 96.6]	89.33 [87.3, 91.1]	71.02 [68.2, 73.9]	69.34 [66.3, 72.3]	85.57 [83.5, 87.5]	80.34 [77.8, 82.8]	84.59 [82.2, 86.8]	84.50 [81.8, 86.9]	78.29 [75.5, 81.0]
	w_{avg} $(\Gamma = \mathbf{R}^\dagger)$	97.90 [97.2, 98.5]	89.40 [87.5, 91.2]	81.75 [79.4, 84.0]	79.23 [76.8, 81.7]	94.06 [92.6, 95.4]	90.64 [89.0, 92.3]	93.92 [92.4, 95.2]	94.65 [93.4, 96.0]	90.17 [88.5, 91.8]
High	w_{μ_θ} $(\Gamma = \mathbf{I})$	**94.88** [93.6, 96.1]	**93.60** [92.2, 94.9]	72.13 [69.2, 75.0]	85.61 [83.2, 87.7]	85.28 [83.1, 87.6]	77.20 [74.4, 80.0]	81.61 [79.0, 84.1]	85.66 [83.4, 88.0]	82.57 [80.1, 85.0]
	w_{μ_θ} $(\Gamma = \mathbf{R}^\dagger)$	93.38 [92.0, 94.6]	88.94 [87.0, 90.8]	**75.72** [73.2, 78.3]	**92.56** [90.9, 94.0]	**87.20** [85.1, 89.2]	**81.86** [79.5, 84.2]	**85.63** [83.5, 87.7]	**86.83** [84.9, 88.7]	85.55 [83.6, 87.7]
	w_{avg} $(\Gamma = \mathbf{I})$	94.08 [92.7, 95.3]	92.13 [90.5, 93.6]	68.60 [65.7, 71.4]	78.57 [75.9, 81.2]	81.71 [79.3, 84.0]	73.19 [70.3, 76.0]	82.91 [80.6, 85.3]	82.96 [80.4, 85.4]	79.32 [76.5, 81.9]
	w_{avg} $(\Gamma = \mathbf{R}^\dagger)$	91.55 [90.0, 93.1]	87.98 [86.0, 89.8]	74.09 [71.0, 76.9]	91.78 [90.2, 93.3]	85.57 [83.3, 87.8]	79.29 [76.6, 81.8]	84.91 [82.7, 87.1]	86.61 [84.5, 88.5]	**85.91** [83.8, 87.9]

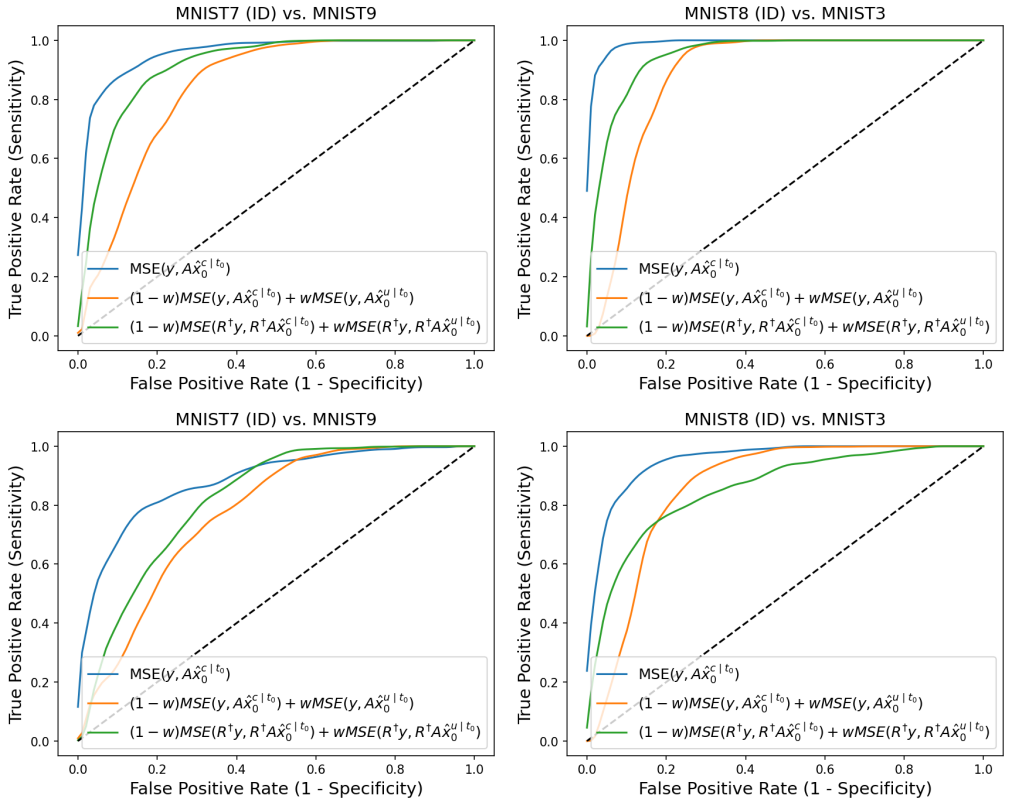
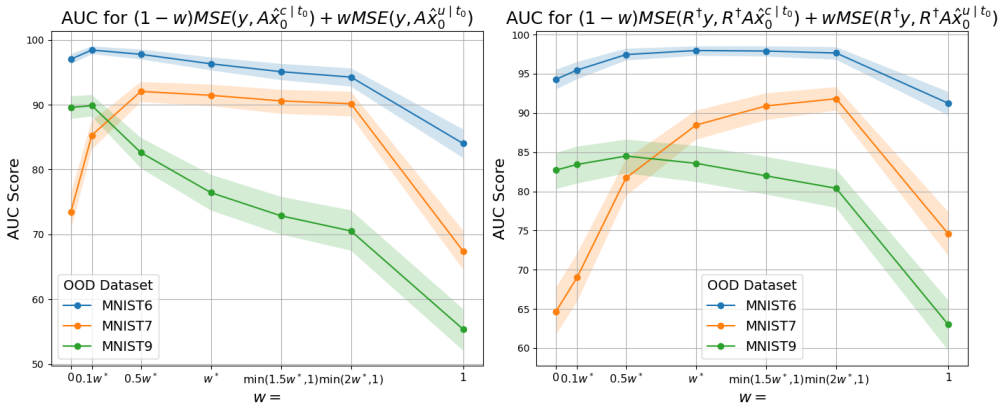
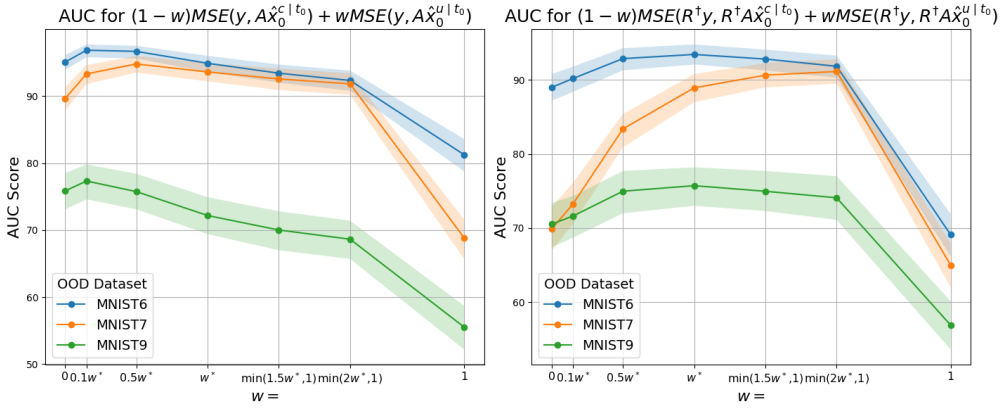


Figure S6: The ROC curves demonstrate the trade-off between TPR (sensitivity) and $1-\text{FPR}$ (specificity) in cases where the proposed weighting scheme resulted in a **lower** AUC compared to using reconstruction error in the projection domain with conditional samples. The top row plots the results for the Moderate Sparsity scenario (with $\{18, 12, 9\}$ projections), whereas the bottom row considers the High Sparsity scenario (with $\{6, 5, 4\}$ projections).



(a) Moderate Sparsity with #proj = $\{18, 12, 9\}$.



(b) High Sparsity with #proj = $\{6, 5, 4\}$.

Figure S7: AUC scores obtained using the proposed weighting scheme, plotted against different scales of w^*, where w^* represents the normalized L^2 distance between the measurement and the forward-projected mean of the training set. When scaled up, values exceeding 1 are capped at 1. The cases $\mathbf{w} = \mathbf{0}$ and $\mathbf{w} = \mathbf{0}$ correspond to conditional and unconditional reconstructions, respectively. This experiment is performed using the model trained on MNIST4, with MNIST6, MNIST7, and MNIST9 considered as OOD datasets. Shaded regions around AUC values indicate confidence bounds, derived from 1000-fold bootstrapping.

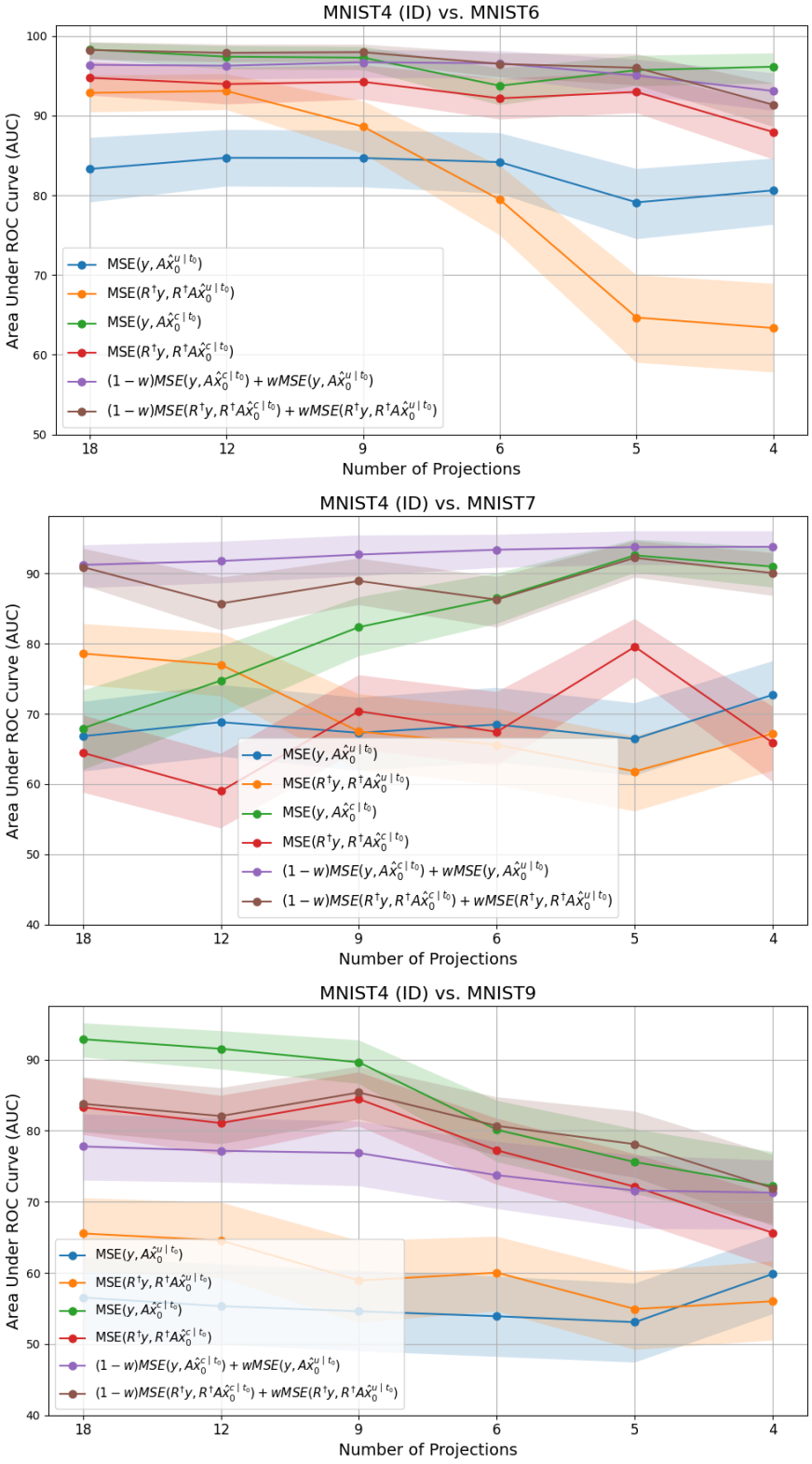


Figure S8: AUC scores across different number of projections, considering six different reconstruction errors discussed in the paper. This experiment is performed using the model trained on MNIST4, with MNIST6, MNIST7, and MNIST9 being the OOD datasets. Shaded regions around AUC values indicate confidence bounds, derived from 1000-fold bootstrapping.

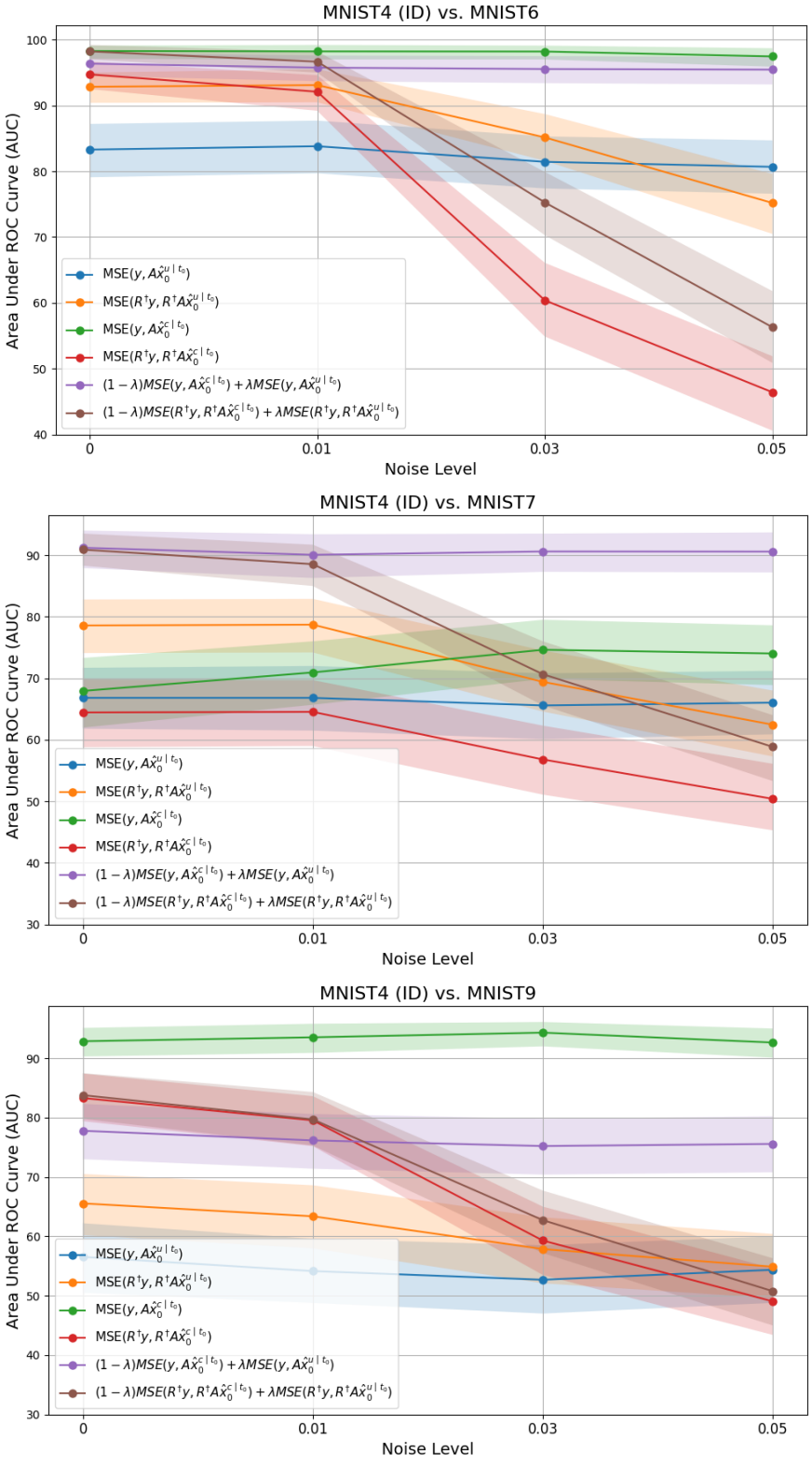


Figure S9: AUC scores across different levels of noise, considering six different reconstruction errors discussed in the paper. This experiment is performed using the model trained on MNIST4, with MNIST6, MNIST7, and MNIST9 being the OOD datasets. The number of projections is kept fixed at 18. Shaded regions around AUC values indicate confidence bounds, derived from 1000-fold bootstrapping.

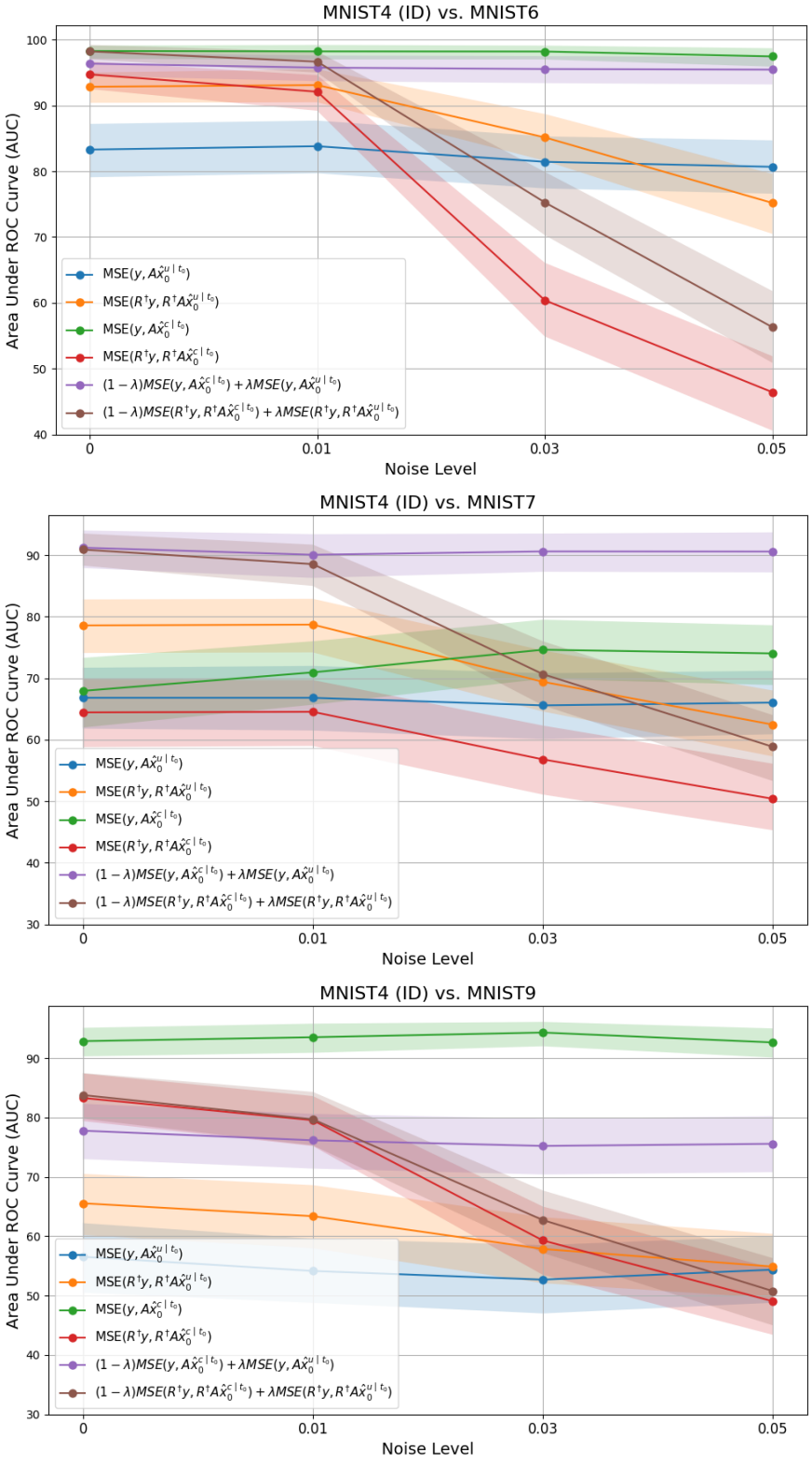


Figure S10: AUC scores across different levels of noise, considering six different reconstruction errors discussed in the paper. This experiment is performed using the model trained on MNIST4, with MNIST6, MNIST7, and MNIST9 being the OOD datasets. The number of projections is kept fixed at 18. Shaded regions around AUC values indicate confidence bounds, derived from 1000-fold bootstrapping.

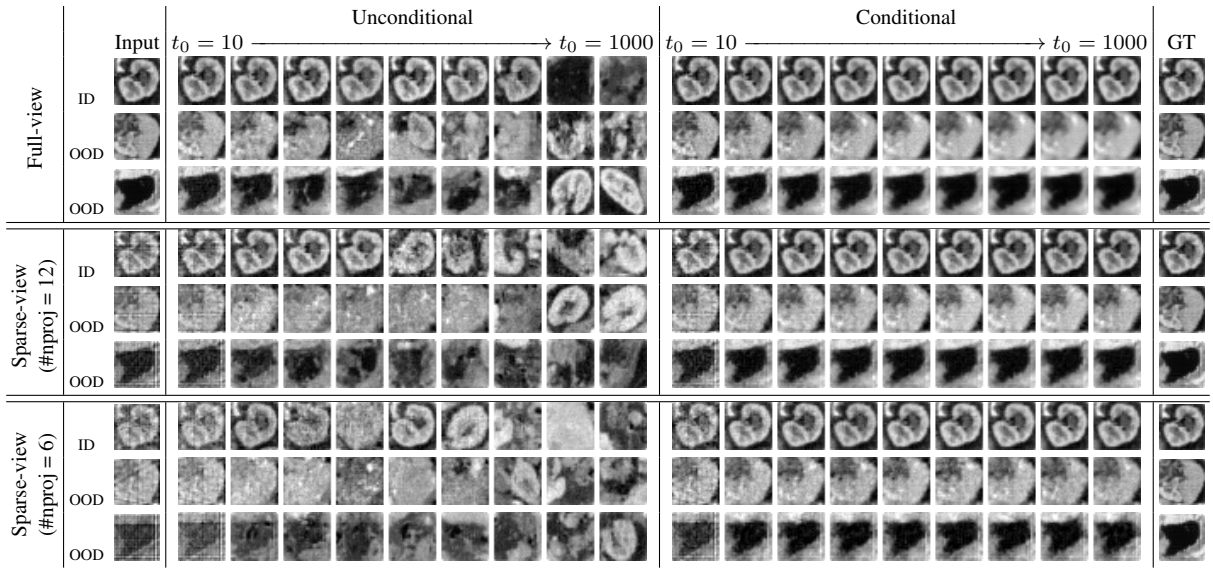


Figure S11: Example unconditional and conditional reconstructions for nine different t_0's, from a model trained on kidneys from MEDMNIST. Rows 1-3 present reconstructions from full-view input images of kidney, liver, and lung respectively, whereas rows 4-6 and 7-9 display reconstructions from sparse reconstructions with 12 and 6 projection angles for the same images, respectively. The images in the final column of the conditional reconstruction section correspond to the estimated CT reconstructions, which can be compared to the ground-truths (GT) shown in the last column of the plot.

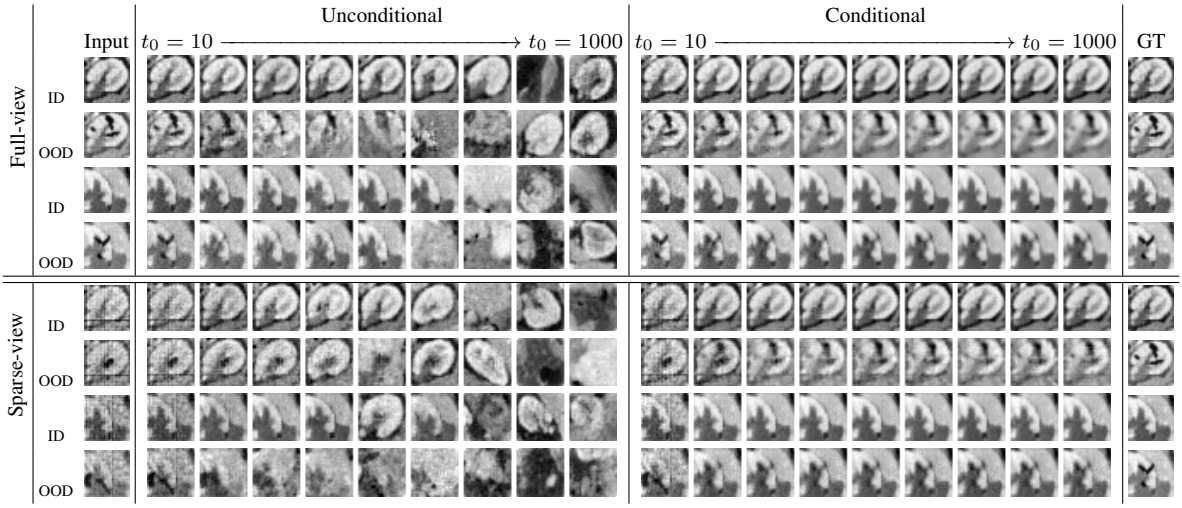


Figure S12: Example unconditional and conditional reconstructions for nine different t_0's, from a model trained on kidney images from MEDMNIST. Rows 1–4 show reconstructions from full-view input images of kidneys (ID) and kidneys with small synthetic anomalies (OOD), while rows 5–8 display reconstructions from FBP mappings of sparse measurements with 9 projection angles for the same images. The final column in the conditional reconstruction section presents the estimated CT reconstructions, which can be compared to the ground-truths (GT) shown in the last column of the plot.