This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available online.

Empirical Characterization of Re-Identification Risk and Diagnostic Utility Under Gerchberg-Saxton Spectral Transformation in Medical Imaging

Anonymous Authors

Featured Application
GS transformation can be applied as a training-free, input-level preprocessing step in medical imaging pipelines prior to dataset sharing, reducing subject-identification risk across multiple adversary classes while preserving diagnostic utility within quantified bounds on brain MRI and chest radiography.

Abstract
Medical images contain anatomical identifiers that persist beyond explicit metadata removal. We empirically characterize how Gerchberg-Saxton (GS) transformation affects subject-identification risk and diagnostic signal across two modalities and four adversary classes. We evaluate transformation impact on OASIS-1 brain MRI and NIH-CXR chest radiography using ResNet18 and DenseNet121, measuring diagnostic utility via subject-level AUC and subject-identification risk under cross-domain and in-domain supervised re-identification (closed-set), embedding-based linkage (open-set), and reconstruction-based inversion using U-Net and diffusion posterior sampling. Cross-domain supervised re-identification collapses to near-chance by GS-10% (OASIS-1 ResNet18: 0.6% [95% CI: 0.0–0.9]; NIH-CXR DenseNet121: 0.6% [0.3–0.8]), from baselines of 86.2% and 72.4%. Reconstruction models achieve high perceptual fidelity (U-Net SSIM: 0.937 [0.932–0.942] at GS-0%) but do not recover subject identity despite direct training exposure to all evaluation subjects. In-domain supervised and embedding-based re-identification decline consistently across masking levels, with modality-dependent magnitudes. Diagnostic AUC reductions remain within 3.6–6.1 percentage points at GS-50%. GS transformation produces a consistent empirical asymmetry between subject-identification risk reduction and diagnostic signal retention across modalities, architectures, and adversary classes. Masking strength p tunes this asymmetry, yielding a range of empirical operating points. Results reflect empirical evaluation under assessed adversary classes and do not constitute formal privacy guarantees.

Keywords: Medical imaging; Subject-identification risk; Re-identification accuracy; Adversarial characterization; Gerchberg-Saxton; Clinical utility-privacy trade-off; Frequency-domain transformation

1. Introduction

The increasing availability of large-scale imaging datasets has accelerated machine learning for healthcare but introduced patient privacy and data-sharing challenges [1, 2]. Unlike structured clinical data, medical images contain detailed anatomical information that can act as implicit identifiers even when explicit personal information is removed [3, 4]. Foundation-model embeddings extracted from medical images can be sufficient to link images to individuals [5], establishing pixel-level re-identification as a demonstrated risk.

Conventional de-identification focuses on metadata removal but does not address identity-related information embedded within image content. Differential privacy limits information exposure through noise injection with documented utility trade-offs in imaging settings [6-8]. Generative and representation learning approaches reduce identity-related features while preserving task-relevant signal but typically

require complex training pipelines that may not generalize across datasets and modalities [9]. Image-domain approaches such as defacing and skull stripping modify content directly but target anatomical visibility rather than learned identity representations, leaving spectral and structural cues that enable re-identification from internal imaging features unaddressed [3-5].

In this work, we investigate how the Gerchberg-Saxton (GS) transformation affects subject-identification risk and diagnostic signal in medical images. GS operates in the frequency domain at the input level, perturbing phase information and masking spectral components without model-specific training or anatomical priors, disrupting fine-grained spatial structure while retaining global statistical properties [10, 11]; the algorithmic specification and structural characterization are provided in Supplementary S1. We evaluate four claims: (1) GS reduces subject-identification risk across multiple adversary classes; (2) the reduction persists against reconstruction adversaries trained on paired RAW-GS data with direct exposure to all evaluation subjects; (3) the reduction persists when the adversary operates within the transformed domain; and (4) diagnostic AUC changes remain within a small, quantified margin. This work characterizes a single transformation across modalities, architectures, and adversary classes; we do not benchmark against alternatives, claim formal privacy properties, or evaluate foundation-model linkage, model inversion, or membership inference. Comparative evaluation and post-training adversaries are identified as follow-up work.

2. Materials and Methods

2.1 Gerchberg-Saxton (GS) Algorithm

The GS algorithm is applied as an input-level transformation [11]. The procedure iteratively alternates between spatial and frequency domains, perturbing phase information and applying stochastic frequency-domain masking. The transformation has two components. Phase perturbation, present at all masking levels, disrupts fine-grained spatial structure by replacing the spectral phase with a randomized initialization [10]. Frequency masking, controlled by masking probability p, progressively removes spectral coefficients.

GS is evaluated at $p \in \{0, 10, 20, 30, 40, 50\}\%$ across utility and subject-identification risk experiments, with reconstruction experiments evaluated at $p \in \{0, 50\}\%$ as the weakest and strongest conditions. The procedure is applied independently per image without labels, subject identifiers, or model-specific information. Transformed images are generated once per experimental condition and reused to ensure observed effects reflect the transformation rather than repeated stochastic sampling. Full algorithmic specification, structural characterization, iteration-convergence and stochastic-stability analyses, and runtime characterization are provided in Supplementary S1–S2.

2.2 Datasets

We evaluate GS on two datasets covering distinct modalities and prediction tasks: structural brain MRI (OASIS-1) and chest radiography (NIH-CXR). OASIS-1 provides T1-weighted structural brain MRI scans; we use it for binary dementia classification (healthy vs. dementia) across 347 subjects [12]. Utility evaluation uses subject-level partitioning across a central slice band. Supervised re-identification uses a slice-level split across 8 non-adjacent slice positions per subject, with all 347 subjects appearing across train, validation, and test partitions via anatomically distinct slice positions from the same acquisition. This constitutes a controlled maximum-sensitivity evaluation of whether identity-bearing structure remains learnable across spatially separated regions of the same volume under GS transformation.

The NIH-CXR dataset consists of frontal radiographs annotated for thoracic pathologies, evaluated as an 11-class multi-label classification task after excluding four classes with insufficient per-subject representation [13]. Utility evaluation uses subject-level partitioning across patients with multiple images.

Supervised re-identification uses an image-level split across 1,739 patients with five or more images, with each patient's images distributed across train, validation, and test partitions, evaluating within-subject identity persistence across separate clinical acquisitions.

For both datasets, utility and embedding-based linkage evaluations enforce strict subject-level assignment with no subject appearing across training and evaluation splits, corresponding to an open-set evaluation. Supervised re-identification splits are slice- or image-level by design, with all subjects present across partitions, corresponding to a closed-set evaluation. Cohort filtering, split sizes, and class lists are provided in Supplementary S3.

2.3 Models

We characterize GS effects using ResNet18 and DenseNet121, selected for their established performance in medical imaging tasks [14] and their structurally distinct connectivity patterns. ResNet18 uses residual skip connections across a compact layer structure [15], while DenseNet121 uses dense connectivity where each layer receives feature maps from all preceding layers across a deeper architecture [16]. Evaluating both lets us assess whether GS-induced changes are architecture-consistent or architecture-specific.

Both models are adapted for single-channel input to accommodate grayscale medical images. Models are trained for binary classification on OASIS-1 (cross-entropy with class weights and label smoothing) and for the 11-class multi-label task on NIH-CXR (binary cross-entropy with per-class positive weighting). Models are trained and evaluated separately for each GS level, with RAW and GS-transformed images never mixed within a training run. Full training details and split-specific evaluation procedures are provided in Supplementary S4-8.

The same utility models are used for embedding-based re-identification evaluation, where penultimate-layer embeddings provide task-aligned representations for similarity-based identity inference. Gallery and probe sets are constructed from the fixed held-out test subjects. Variability across model instantiations therefore reflects differences in utility model training rather than evaluation-population differences, enabling direct subject-level comparison. Performance is evaluated at the subject level with predictions aggregated across images from the same individual.

2.4 Experimental Conditions

Figure 1 illustrates the experimental framework. We characterize subject-identification risk under four measurement conditions that vary in evaluation domain and assumed data access.

Cross-domain supervised re-identification. This condition measures whether identity-related features learned from GS-transformed representations transfer to RAW untransformed images. A closed-set classifier is trained on GS-transformed images and evaluated on RAW images using the closed-set split defined in 2.2.

Reconstruction-based inversion. This condition measures whether RAW images can be reconstructed from GS-transformed images, and whether the reconstructions retain identity-bearing structure. Two paradigms are evaluated: a U-Net [17] trained to minimize pixel-wise reconstruction error, and a diffusion posterior sampling model [18] under pretrained and fine-tuned conditions. Both models are trained with direct exposure to the RAW anatomical appearance of all evaluation subjects via paired RAW-GS data, with only specific test slices withheld. Separate models are trained at GS-0% and GS-50%. Reconstructed outputs are evaluated using the same re-identification classifiers and same splits used in supervised re-identification experiments.

In-domain supervised re-identification. Even if GS prevents identity transfer to RAW images, identity-bearing structure may persist within the transformed representation itself. This condition measures how much subject-identification risk remains learnable entirely within the GS domain, independent of any domain boundary crossing. A closed-set classifier trains and evaluates on GS- transfor-

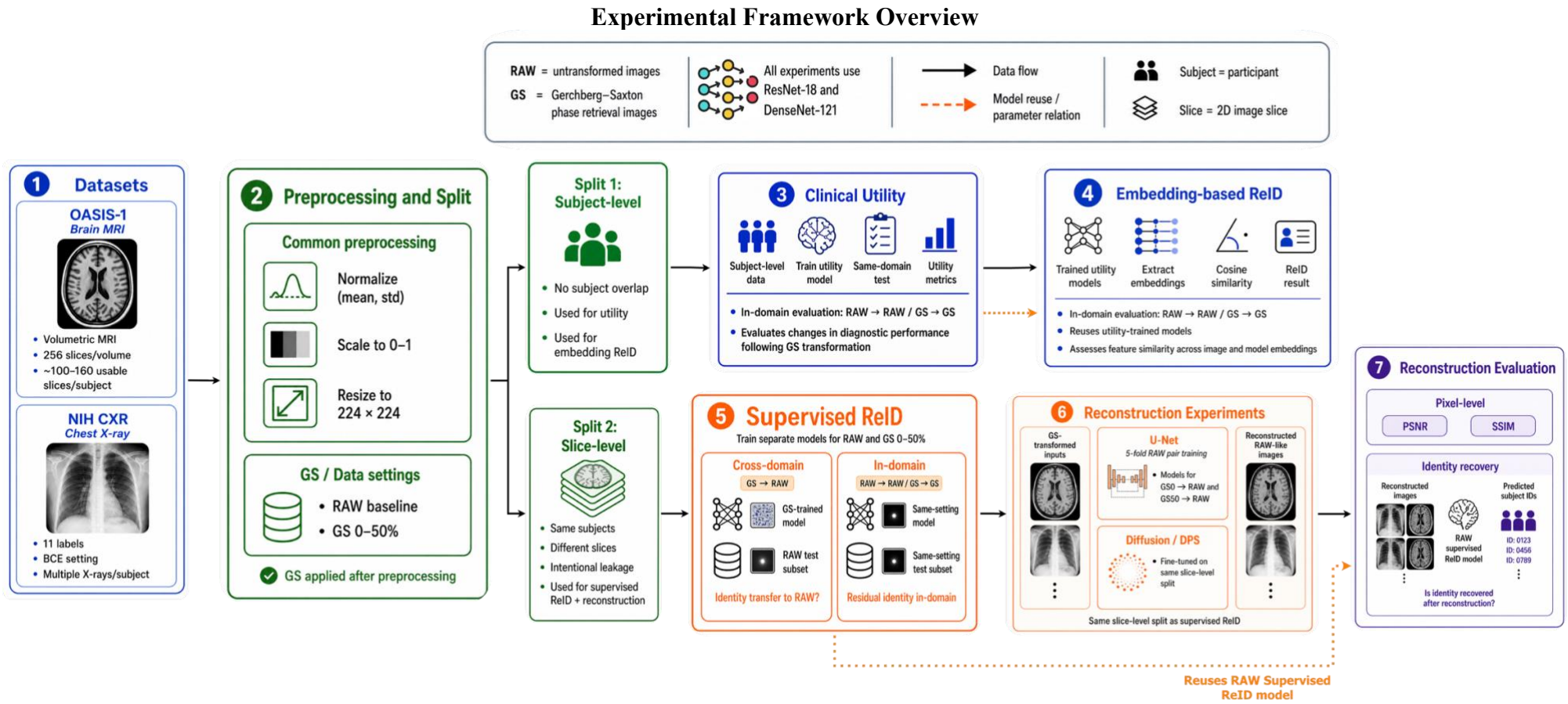


Figure 1. Datasets are preprocessed and split at subject level for utility and embedding-based re-identification evaluation, and at slice level for supervised re-identification and reconstruction-based inversion. GS transformation is applied independently per image after preprocessing across masking levels $p \in \{0\%, 10\%, 20\%, 30\%, 40\%, 50\%\}$. Utility models are trained and evaluated under matched transformation conditions; the same trained models are reused for embedding-based re-identification evaluation via penultimate-layer cosine similarity matching. Cross-domain and in-domain supervised re-identification classifiers are trained separately at each GS level using the slice-level split. Reconstruction-based inversion evaluates whether GS-transformed images can be inverted to recover identity-bearing structure using U-Net and diffusion posterior sampling, with reconstructed outputs evaluated by the same RAW-trained re-identification classifier used across all supervised re-identification conditions. Solid arrows indicate data flow; dashed orange arrows indicate model reuse.

med images at the same masking level, establishing an upper bound on within-domain re-identification accuracy.

Embedding-based re-identification. This open-set condition measures whether penultimate-layer embeddings from utility-trained models retain subject-identification risk under GS transformation. Embeddings are extracted from the average pooling layer of ResNet18 (512 dimensions) and the feature pooling layer of DenseNet121 (1,024 dimensions), L2-normalized, and matched using cosine similarity against a GS-transformed reference gallery. Identity inference is evaluated as Rank-1 accuracy. Gallery size k denotes the number of reference images per subject and is varied at $k = 1$ and $k = 5$ under matched transformation conditions.

Together, these four experiments correspond to distinct re-identification questions. The canonical record of subject identity in this study is the RAW image, since this is the form in which anatomical and biometric features are originally acquired and the form in which any reference data held by an adversary is most likely represented [3-5]. Cross-domain re-identification therefore measures the primary question of interest: whether features learned from a transformed GS image can be matched back to a RAW reference. Reconstruction-based inversion measures the same RAW-recovery question under adversaries that has paired (RAW, GS) supervision and attempts reconstructing RAW versions of the GS transformed images. In contrast, embedding-based and in-domain re-identification both operate under matched-domain conditions (RAW-to-RAW or GS-to-GS) and characterize how much identity-bearing structure remains intrinsically present in GS representations. The difference is that embedding-based is an open-set evaluation using penultimate features of the utility model, while in-domain is a closed-set supervised evaluation. Together they bound residual within-domain learnability, distinct from the GS-to-RAW linkage assessed by the other two conditions.

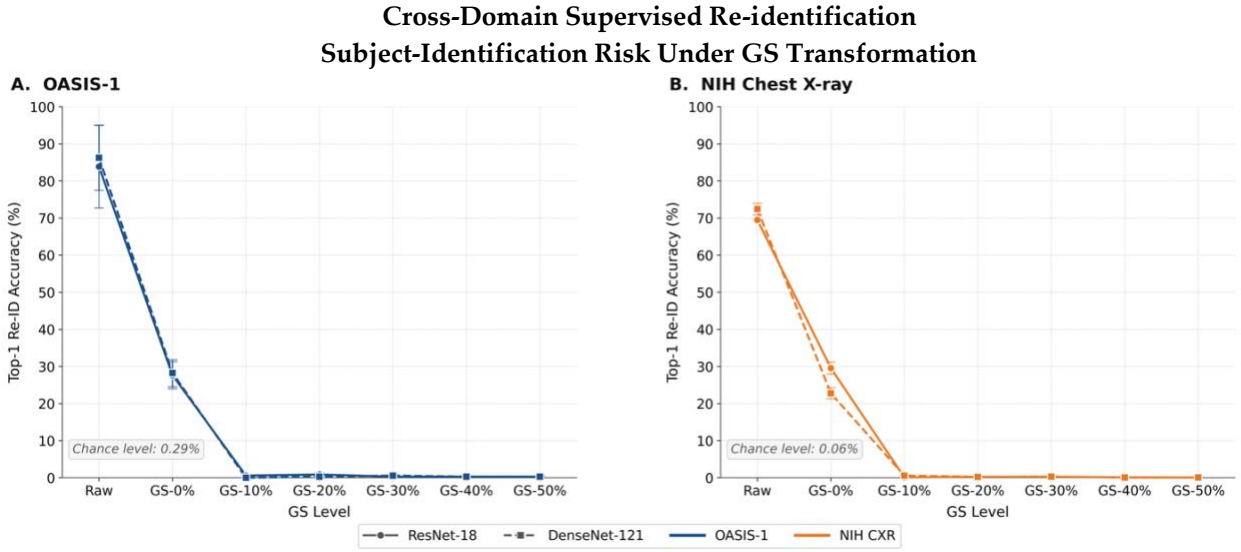


Figure 2 — Top-1 re-identification accuracy under cross-domain supervised re-identification across GS masking levels for OASIS-1 (A, 347 subjects; chance = 0.29%) and NIH-CXR (B, 1,739 subjects; chance = 0.06%). Classifiers are trained on GS-transformed images and evaluated on RAW images. Error bars represent 95% confidence intervals.

2.5 Evaluation Metrics

Diagnostic utility is assessed using subject-level AUC under matched transformation conditions (RAW-to-RAW, GS-to-GS), with predictions aggregated across images from the same subject. Mean values and 95% confidence intervals are computed from 5-fold cross-validation t-intervals (df=4) for OASIS-1 and

subject-level bootstrap resampling (B=2000) for NIH-CXR. Procedures for reconstruction outputs and detailed formulations are provided in Supplementary S7 and S8.

Reconstruction fidelity is measured using structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR). Re-identification accuracy from reconstructed images is using the same re-identification classifiers and splits used across all supervised re-identification conditions, with CIs following the dataset-specific procedures stated above. SSIM and PSNR confidence intervals are derived from image-level bootstrap resampling (B=2000).

3. Results

3.1 GS Alters Identity-Related Feature Transferability: Cross-Domain Supervised Re-Identification

Cross-domain supervised re-identification accuracies across both datasets and architectures are summarized in Table 1 and Figure 2.

At RAW baseline, re-identification accuracy is high across both datasets and architectures, confirming that RAW medical images carry strong identity-related signal. After GS transformations, phase perturbation alone at GS-0% drives a substantial initial drop (~56.2-60.0% on OASIS-1, ~39.9-49.6% on NIH-CXR from RAW) without any frequency masking on both datasets and architectures. This reflects the random-phase modulation applied per image (Supplementary S1.5). By GS-10%, accuracy collapses to near-chance across all conditions (range: 0.0-0.6%) and remains there through GS-50%.

This pattern indicates that GS transformation severs the correspondence between transformed-domain features and RAW-image identity, rather than reflecting general degradation of model performance. Even minimal frequency masking (GS-10%) is sufficient to sever this correspondence, and near-chance performance is stable through GS-50%, indicating that additional masking does not reduce accuracy further beyond this point. This establishes the first component of the utility-identity asymmetry; the joint view with diagnostic utility appears in Section 3.5.

GS Level	OASIS-1		NIH CXR	
	ResNet-18 [95% CI]	DenseNet-121 [95% CI]	ResNet-18 [95% CI]	DenseNet-121 [95% CI]
RAW	83.9 [72.7, 95.0]	86.2 [77.5, 95.0]	69.5 [69.2, 69.8]	72.4 [70.8, 74.0]
GS-0%	27.7 [24.0, 31.3]	28.2 [24.5, 31.8]	29.6 [27.9, 31.2]	22.8 [21.2, 24.3]
GS-10%	0.6 [0.0, 0.9]	0.0 [0.0, 0.0]	0.1 [0.0, 0.2]	0.6 [0.3, 0.8]
GS-20%	0.9 [0.0, 1.4]	0.3 [0.0, 0.5]	0.2 [0.1, 0.4]	0.2 [0.1, 0.4]
GS-30%	0.3 [0.0, 0.5]	0.6 [0.0, 0.9]	0.2 [0.1, 0.4]	0.3 [0.1, 0.5]
GS-40%	0.3 [0.0, 0.5]	0.3 [0.0, 0.5]	0.1 [0.0, 0.2]	0.1 [0.0, 0.1]
GS-50%	0.3 [0.0, 0.5]	0.3 [0.0, 0.5]	0.1 [0.0, 0.1]	0.1 [0.0, 0.1]

Table 1. Cross-domain supervised re-identification Top-1 accuracy (%) with 95% confidence intervals across GS masking levels. CIs from 5-fold cross-validation t-intervals (OASIS-1) and subject-level bootstrap (NIH-CXR). Chance level: 0.29% (OASIS-1), 0.06% (NIH-CXR).

3.2 Subject-Identification Risk Declines Under Same-Domain Conditions

Same-domain re-identification results are summarized in Table 2 and Figure 3. The decline is more gradual than the cross-domain collapse in 3.1: but expected considering the classifier operates entirely within the transformed domain and does not bridge the RAW-GS boundary.

In-domain supervised re-identification. Re-identification accuracy declines consistently with increasing masking across both datasets and architectures, though the pattern and magnitude differ

substantially between modalities. On NIH-CXR, the decline is steep and consistent. RAW baselines (69.5% R18, 72.4% D121) drop to 42.1% / 39.9% at GS-10% and 13.4% / 15.4% at GS-50% (CIs in Table 2). By GS-50%, accuracy is substantially reduced, indicating that GS disrupts identity-bearing structure in 2D frontal radiographs even within the transformed domain. On OASIS-1, the decline is more gradual. RAW baselines (83.9% R18, 86.2% D121) drop to 72.2% / 75.5% at GS-10% and 52.8% / 56.3% at GS-50% (CIs in Table 2). Despite the consistent decline, re-identification accuracy remains substantially above chance at GS-50% under both architectures. This residual learnable structure exists entirely within the transformed domain,

Dataset	GS Level	ResNet-18			DenseNet-121		
		Oracle Top-1	FE k=1	FE k=5	Oracle Top-1	FE k=1	FE k=5
OASIS-1	RAW	83.9 [72.7, 95.0]	54.2 [42.2, 66.1]	81.8 [67.9, 95.7]	86.2 [77.5, 95.0]	51.6 [45.6, 57.5]	79.8 [70.5, 89.2]
	GS-0%	80.4 [68.5, 92.2]	46.1 [38.6, 53.6]	75.2 [65.0, 85.4]	84.5 [76.4, 92.7]	45.0 [39.9, 50.0]	73.2 [64.2, 82.2]
	GS-10%	72.2 [58.1, 86.2]	34.0 [28.8, 39.2]	60.2 [56.8, 63.7]	75.5 [65.5, 85.6]	33.4 [29.0, 37.8]	62.5 [57.2, 67.8]
	GS-20%	68.4 [53.8, 83.0]	28.3 [21.0, 35.5]	52.2 [40.0, 64.3]	66.4 [53.1, 79.6]	27.4 [16.4, 38.4]	56.8 [49.7, 63.9]
	GS-30%	63.7 [49.3, 78.0]	23.6 [19.0, 28.3]	53.6 [43.7, 63.6]	63.3 [48.2, 78.5]	25.6 [21.9, 29.3]	45.2 [38.4, 52.1]
	GS-40%	59.1 [46.9, 71.3]	16.1 [13.1, 19.2]	43.8 [30.9, 56.7]	60.0 [44.5, 75.6]	17.6 [13.5, 21.7]	39.8 [31.7, 47.8]
	GS-50%	52.8 [38.7, 66.8]	18.2 [13.5, 22.8]	37.7 [32.1, 43.4]	56.3 [39.1, 73.5]	11.8 [6.9, 16.7]	29.7 [21.9, 37.4]
NIH CXR	RAW	69.5 [69.2, 69.8]	19.8 [17.5, 22.2]	22.4 [19.9, 24.9]	72.4 [70.8, 74.0]	26.9 [24.2, 29.5]	26.4 [23.8, 29.2]
	GS-0%	62.5 [60.8, 64.2]	13.9 [11.8, 16.0]	14.9 [12.8, 17.1]	63.6 [62.6, 64.7]	16.3 [14.2, 18.6]	17.0 [14.8, 19.3]
	GS-10%	42.1 [40.8, 43.5]	4.4 [3.3, 5.8]	6.5 [5.2, 7.9]	39.9 [38.8, 40.9]	9.2 [7.5, 11.0]	10.7 [8.8, 12.5]
	GS-20%	32.1 [31.4, 32.8]	2.8 [1.9, 3.8]	4.7 [3.5, 6.0]	31.1 [30.2, 31.9]	6.6 [5.2, 8.1]	6.7 [5.3, 8.2]
	GS-30%	24.6 [23.1, 26.2]	1.8 [1.0, 2.6]	3.2 [2.2, 4.2]	25.2 [24.2, 26.3]	3.6 [2.5, 4.7]	3.8 [2.7, 4.9]
	GS-40%	18.3 [17.0, 19.7]	1.8 [1.0, 2.6]	2.8 [1.9, 3.9]	20.3 [19.1, 21.6]	3.1 [2.1, 4.2]	4.9 [3.7, 6.2]
	GS-50%	13.4 [11.3, 15.6]	1.4 [0.8, 2.2]	2.8 [1.9, 3.9]	15.4 [14.0, 16.8]	3.0 [2.0, 4.2]	4.7 [3.5, 6.0]

Table 2. In-domain supervised re-identification Top-1 accuracy (%) and embedding-based re-identification Rank-1 accuracy (%) with 95% confidence intervals across GS masking levels. RAW baseline rows are identical across in-domain and cross-domain supervised re-identification conditions, as both reflect RAW-to-RAW evaluation. CIs from 5-fold cross-validation (OASIS-1) and subject-level bootstrap (NIH-CXR). k = gallery size.

and it does not transfer back to RAW images as established in 3.1; thus, these results represent within-domain learnability bounds what an adversary can extract from transformed GS images alone rather than indicating recoverable identity. Furthermore, the steeper decline on NIH-CXR compared to OASIS-1 is notable. Our post-hoc interpretation of this pattern is that modality-dependent imaging characteristics drive the difference in residual learnability. Frontal chest radiographs acquired across different clinical timepoints carry less spatially coherent cross-image identity structure, whereas volumetric brain MRI encodes consistent identity-bearing features across anatomically proximate slices within a single acquisition, making residual within-domain learnability more persistent.

Embedding-based re-identification. Accuracy declines consistently across both datasets and is more sensitive to GS than in-domain supervised re-identification. On NIH-CXR, accuracy approaches zero by GS-20% to GS-50% under both gallery sizes. On OASIS-1, the decline is more gradual, with larger galleries amplifying the effect at higher masking levels (RAW → GS-50%: **k=1:** *51.6%-54.2%* → *11.8%-18.2%*; **k=5:** *79.8%-81.8%* → *29.7%-37.7%*). Despite this steeper overall decline, OASIS-1 retains higher absolute re-identification accuracy than NIH-CXR across all masking levels and gallery sizes. This is consistent with the stronger residual identity signal observed under in-domain supervised re-identification. The convergence across two distinct experimental designs (2.4) further supports the post-hoc modality interpretation.

3.3 Reconstruction-Based Inversion Does Not Recover Subject Identity

Reconstruction-based inversion results are summarized in Table 3 and Figure 4. Diffusion-based reconstruction is evaluated on OASIS-1 only. OASIS-1 retains higher residual identity signal under direct re-identification (Sections 3.1–3.2), making it the more conservative test setting [19]. Evaluating the stronger reconstruction adversary on the more resistant modality establishes the upper-bound condition for identity recovery.

In-Domain Supervised Re-identification and Embedding-Based Re-identification Under GS Transformation

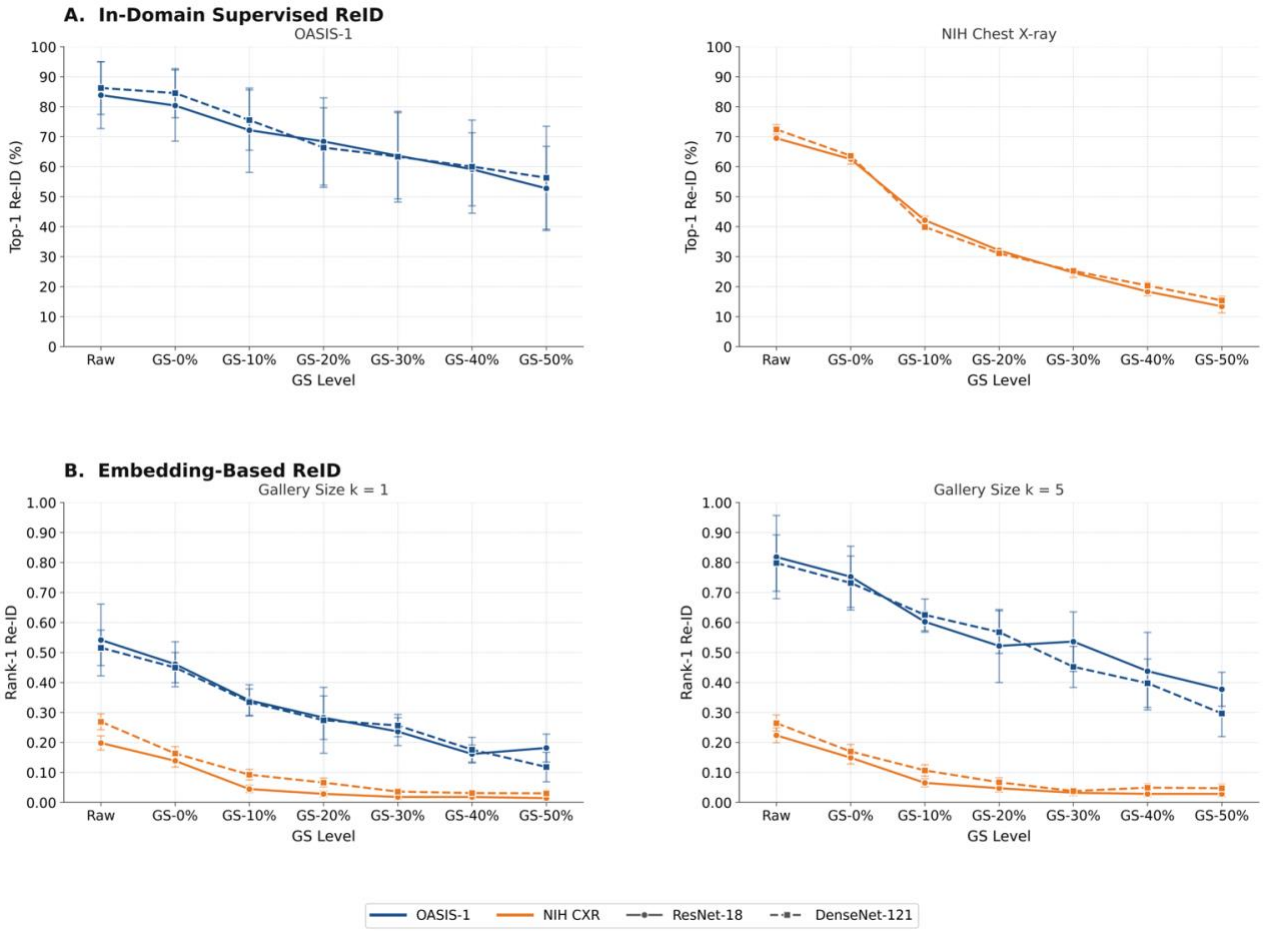


Figure 3 — Re-identification accuracy under same-domain conditions across GS masking levels. (A) In-domain supervised re-identification Top-1 accuracy for OASIS-1 (left) and NIH-CXR (right), where the classifier trains and evaluates within the same GS-transformed domain. (B) Embedding-based re-identification Rank-1 accuracy at gallery size k=1 (left) and k=5 (right), with both datasets represented on each plot. Both conditions show consistent decline with increasing masking strength.

U-Net reconstructions achieve high perceptual fidelity at GS-0% (**SSIM:** 0.937 *OASIS-1*; *0.919* NIH-CXR) and remain structurally coherent at GS-50% despite visible smoothing. This finding is notable given that both reconstruction models were trained with direct exposure to the RAW anatomical appearance of all evaluation subjects via paired RAW-GS training data, representing a strong upper-bound condition for reconstruction-based identity recovery. Despite near-perfect perceptual reconstruction at GS-0%, re-identification accuracy declines substantially from RAW baselines across both datasets and architectures. This observation indicates that the random-phase initialization at GS-0% (Supplementary S1.5) disrupts identity-bearing structure in a way that persists through reconstruction even when global image

appearance is largely recovered. At GS-50%, re-identification collapses to near-chance across both datasets and architectures despite this same training condition.

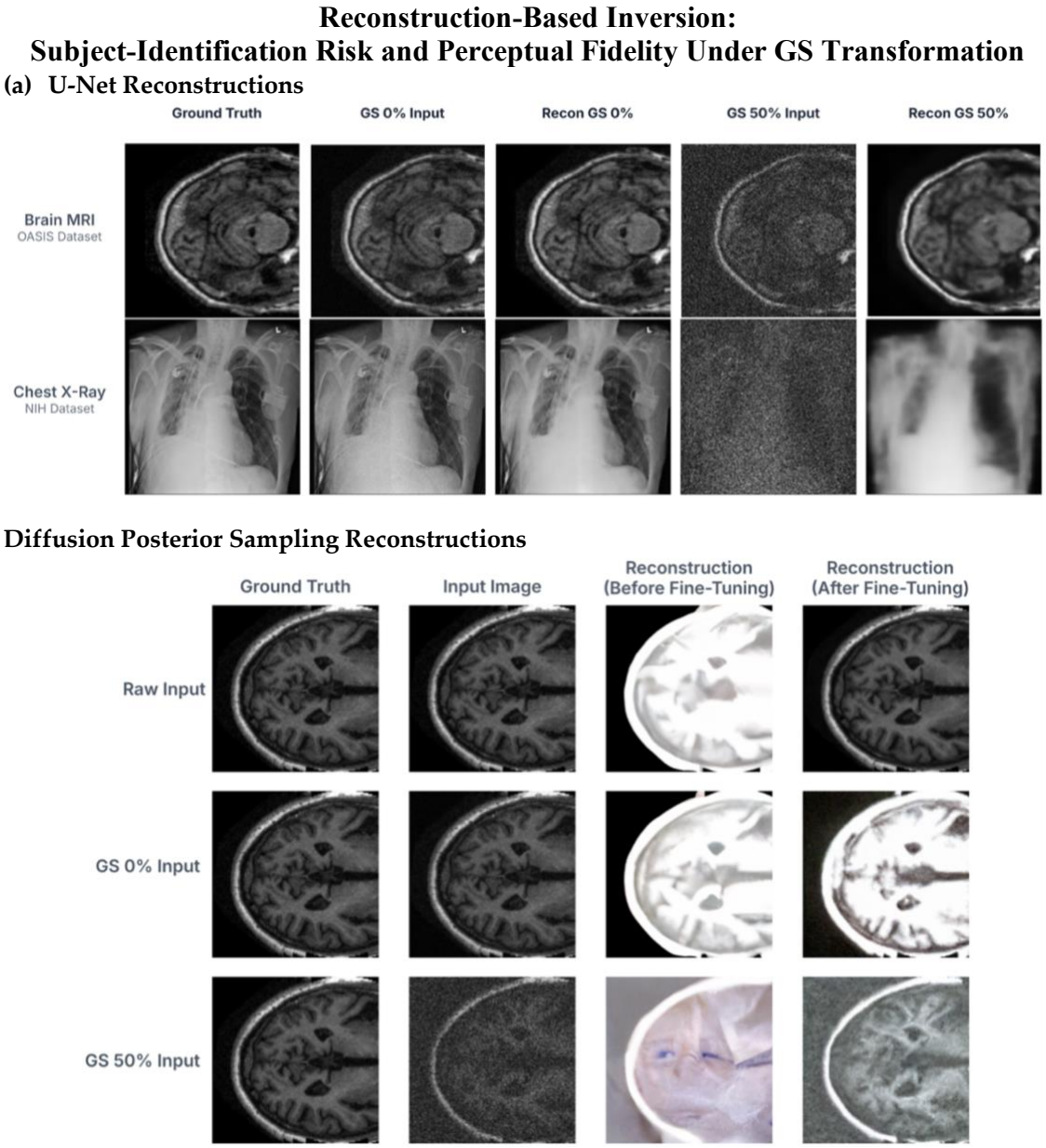


Figure 4 — Representative reconstructions at GS-0% and GS-50% for OASIS-1 and NIH-CXR. (A) U-Net reconstructions remain visually plausible across masking levels but exhibit smoothing and loss of fine structural detail at GS-50%. Despite high perceptual fidelity at GS-0%, re-identification accuracy declines substantially from RAW baselines. (B) DPS reconstructions on OASIS-1 before and after fine-tuning. The pretrained model captures outer anatomical boundaries at GS-0% but fails to recover internal brain structure; at GS-50%, prior-domain structural features become visible within the captured outer shape under severe spectral information loss. Fine-tuned outputs are anatomically coherent but subject-nonspecific, reflecting successful domain adaptation without recovery of identity-bearing structure. Perceptual plausibility does not imply restored re-identification accuracy; quantitative results are reported in Table 3.

Diffusion-based reconstructions reinforce this dissociation and reveal additional nuance. Before fine-tuning, the model's behavior differs qualitatively across masking conditions. At GS-0%, outer anatomical boundaries are captured but fine-grained internal brain structure is not recovered, indicating that phase

perturbation alone is sufficient to prevent the model from resolving subject-specific detail. At GS-50%, severe spectral information loss shifts the model toward its FFHQ pretraining prior, producing visible non-brain structural features within the captured outer shape (Figure 4B). This condition-dependent prior reliance motivates domain-specific fine-tuning as a necessary step for meaningful reconstruction evaluation.

Following fine-tuning, the model produces anatomically coherent brain reconstructions at both masking levels, confirming successful domain adaptation. Re-identification accuracy nonetheless remains near-zero at both conditions, consistent with the U-Net results. At GS-0%, phase perturbation preserves more spectral information but introduces structural misalignment, yielding lower SSIM (0.399). At GS-50%, despite a higher SSIM (0.710), severe spectral loss drives the model toward population-level anatomical priors rather than subject-specific structure, consistent with the FFHQ-drift observed before fine-tuning. Notably, the fine-tuned model was trained on GS-RAW pairs of the exact test subjects. That re-identification nonetheless remains near-zero further supports that GS transformation mitigates identity recovery even under this upper-bound condition.

Dataset	Method	GS Level	SSIM [95% CI]	PSNR dB [95% CI]	Re-ID [95% CI] (R18)	Re-ID [95% CI] (D121)
OASIS-1	RAW		—	—	83.9 [72.7, 95.0]	86.2 [77.5, 95.0]
	U-Net	GS 0%	0.937 [0.932, 0.942]	33.44 [32.94, 33.92]	24.15 [18.36, 30.43]	45.89 [39.13, 52.66]
		GS 50%	0.719 [0.713, 0.724]	26.50 [26.29, 26.70]	2.90 [0.97, 5.31]	1.93 [0.48, 3.86]
	DPS	GS 0%	0.399 [0.381, 0.420]	5.94 [5.80, 6.11]	2.00 [0.00, 6.00]	0.00 [0.00, 0.00]
		GS 50%	0.710 [0.694, 0.725]	12.09 [11.72, 12.43]	2.00 [0.00, 6.00]	0.00 [0.00, 0.00]
NIH CXR	RAW		—	—	69.5 [69.2, 69.8]	72.4 [70.8, 74.0]
	U-Net	GS 0%	0.919 [0.917, 0.921]	30.46 [30.28, 30.65]	22.71 [21.28, 24.17]	14.50 [13.27, 15.77]
		GS 50%	0.737 [0.734, 0.739]	22.30 [22.22, 22.39]	0.16 [0.03, 0.32]	0.19 [0.06, 0.36]

Table 3. Perceptual reconstruction quality (SSIM, PSNR) and re-identification Top-1 accuracy at GS-0% and GS-50%. RAW row serves as the upper-bound reference. Re-identification CIs from 5-fold cross-validation (OASIS-1) and subject-level bootstrap (NIH-CXR); SSIM and PSNR CIs from image-level bootstrap. DPS evaluated on OASIS-1 only.

3.4 Diagnostic Utility Remains Stable Under GS Transformation

Diagnostic utility results are summarized in Table 4 and Figure 5. Across both datasets, AUC remains stable at low masking levels and declines gradually with increasing masking strength. On OASIS-1, performance at GS-0% and GS-10% is within CI overlap of the RAW baseline, indicating that phase perturbation alone does not meaningfully reduce diagnostic signal. From GS-20% onward, decline becomes more consistent, reaching 0.843 (R18) and 0.847 (D121) at GS-50%, compared to RAW baselines of 0.875 and 0.902. On NIH-CXR, decline is more consistent across masking levels, with GS-50% AUC reaching 0.701 (R18) and 0.700 (D121), compared to RAW baselines of 0.743 for both architectures (CIs in Table 4).

This gradual decline contrasts with the steep subject-identification risk declines in Sections 3.1-3.3. In contrast to the modality-dependent re-identification differences in 3.2, diagnostic utility decline remains comparably bounded across both datasets, suggesting GS affects identity-bearing and task-relevant signal through distinct modality-sensitivities. At GS-50%, where cross-domain re-identification has collapsed to near-chance and reconstruction-based inversion fails to recover identity, models retain diagnostic performance. This asymmetry is the central empirical finding of this work and is examined jointly in 3.5.

GS Level	OASIS-1		NIH CXR	
	AUC [95% CI] (R18)	AUC [95% CI] (D121)	AUC [95% CI] (R18)	AUC [95% CI] (D121)
RAW	0.875 [0.841, 0.908]	0.902 [0.875, 0.930]	0.743 [0.731, 0.756]	0.743 [0.730, 0.756]
GS-0%	0.879 [0.871, 0.888]	0.893 [0.846, 0.939]	0.734 [0.721, 0.747]	0.751 [0.738, 0.764]
GS-10%	0.880 [0.840, 0.921]	0.881 [0.827, 0.934]	0.726 [0.712, 0.739]	0.738 [0.724, 0.751]
GS-20%	0.864 [0.829, 0.899]	0.879 [0.843, 0.915]	0.723 [0.710, 0.737]	0.736 [0.722, 0.749]
GS-30%	0.854 [0.825, 0.883]	0.864 [0.827, 0.900]	0.721 [0.708, 0.733]	0.732 [0.719, 0.745]
GS-40%	0.839 [0.809, 0.868]	0.867 [0.816, 0.918]	0.714 [0.700, 0.727]	0.724 [0.710, 0.738]
GS-50%	0.843 [0.788, 0.898]	0.847 [0.802, 0.892]	0.701 [0.687, 0.715]	0.700 [0.687, 0.713]

Table 4. Subject-level AUC under matched GS training and evaluation conditions. CIs from 5-fold cross-validation (OASIS-1) and subject-level bootstrap (NIH-CXR).

The AUC values reported here are within the range of published baselines on these datasets and serve as within-paper reference points rather than deployment benchmarks. The asymmetry characterized in this work is a relative finding under matched training and evaluation conditions: utility decline is small in comparison to identification-risk decline. Whether the residual utility at any specific GS level is clinically sufficient depends on task-specific operating-point requirements (sensitivity at fixed specificity, calibration, decision thresholds) that are out of scope for this characterization.

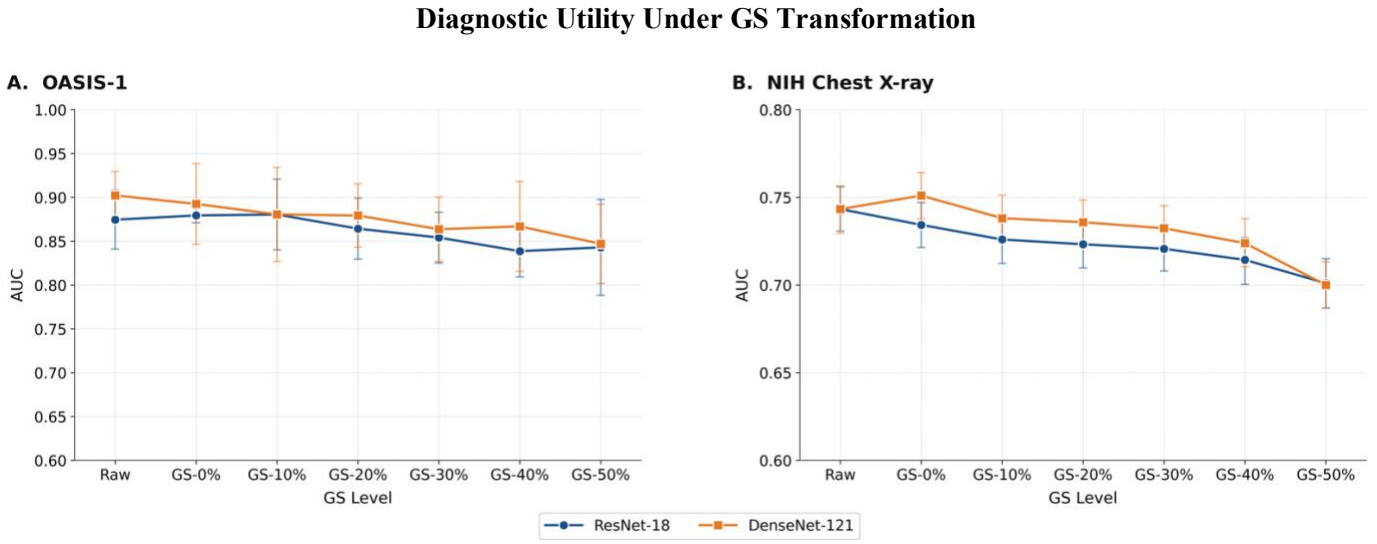


Figure 5 — Subject-level AUC as a function of GS masking strength for OASIS-1 (A) and NIH-CXR (B) using ResNet18 and DenseNet121. Performance declines gradually with increasing masking, with AUC reductions remaining within 3.2-5.5% of the RAW baseline at GS-50% across both datasets and architectures. Error bars represent 95% confidence intervals (5-fold CV for OASIS-1, subject-level bootstrap for NIH).

3.5 Utility-Identity Trade-off

Figure 6 summarizes the joint behavior of diagnostic utility and subject-identification risk across GS masking levels. Across all conditions and datasets, subject-identification risk declines substantially with increasing masking while diagnostic utility changes remain comparatively small. Under cross-domain conditions (Panel A), re-identification accuracy drops by more than 99% points from RAW to GS-10% across all conditions, where the chance floor is reached. Utility changes in Panel C remain within 6.12% points at GS-50% across all conditions, with some architectures showing negligible change at low masking

levels. This asymmetry is consistent across datasets and architectures, indicating that GS transformation affects identity-bearing features more than diagnostic image content.

The in-domain panel (Panel B) provides a complementary view. As shown in 3.2, in-domain re-identification declines more gradually than cross-domain, but the magnitude of decline still substantially exceeds utility loss at every masking level. Embedding-based re-identification declines more steeply than in-domain supervised re-identification, approaching near-zero on NIH-CXR by GS-20%. Embedding-based retrieval is reported only under matched-domain conditions; cross-domain comparison with RAW probes would reflect identity in the RAW input rather than the transformed representation.

Across the conditions evaluated, GS-0% (phase perturbation alone) constitutes a distinct lower-strength operating point, reducing cross-domain re-identification by 57.5% or more relative to RAW while preserving diagnostic AUC within CI overlap on OASIS-1. At GS-10%, cross-domain supervised re-identification collapses to near-chance and in-domain supervised re-identification declines substantially, while diagnostic AUC remains within 2.4% of the RAW baseline across all conditions. Beyond GS-10%, re-identification remains near-chance across all conditions while diagnostic AUC reductions accumulate modestly, with the largest utility cost observed at GS-50% remaining within 6.12%. Together, these results characterize a range of empirical operating points in which moderate to strong GS masking consistently reduces subject-identification risk while diagnostic utility reductions remain bounded.

4. Discussion

The central finding of this work is an empirical asymmetry: subject-identification risk declines sharply under GS masking while diagnostic utility decreases only modestly, and this pattern holds consistently across two datasets, two architectures, and four adversary classes. The result indicates that GS affects identity-bearing features more than task-relevant diagnostic signal, aligning with the broader concern that medical images inherently encode identifiable information beyond explicit metadata [3-5, 20].

Among the four adversary classes, cross-domain supervised re-identification carries the primary RAW-linkage question: whether features learned from GS-transformed representations can be matched back to a RAW reference. Re-identification accuracy collapses to near-chance by GS-10% across both datasets and architectures, indicating that this transfer does not occur. Residual re-identification observed under in-domain conditions therefore reflects learnable identity structure that remains within the transformed domain but does not project back to RAW images, bounding adversarial risk to within-domain learnability rather than RAW-image linkage.

The modality-dependent pattern emerges under both in-domain supervised and embedding-based re-identification. The rate and magnitude of decline differ substantially between modalities, suggesting that volumetric brain MRI encodes more resistant identity-bearing structure under GS transformation than 2D frontal radiographs. In contrast, diagnostic utility decline remains comparably bounded across both modalities, indicating that GS transformation affects identity-bearing and diagnostic features through distinct mechanisms that are not equally modality-dependent.

Another key finding is the dissociation between reconstruction fidelity and identity recovery. U-Net and DPS reconstructions remain visually plausible while re-identification accuracy is substantially reduced. Both reconstruction models were trained with direct exposure to the RAW anatomical appearance of all evaluation subjects, representing a strong upper-bound condition for reconstruction-based identity recovery. The failure to recover re-identification accuracy under this design reflects a genuine inability to restore identity-bearing structure from GS-transformed inputs rather than a limitation of model capacity. This pattern is consistent with established perspectives in inverse problems, where reconstruction models recover plausible solutions guided by learned priors rather than the original signal [21, 22]. Perceptual

similarity alone is therefore not a reliable indicator of subject-identification risk under the evaluated reconstruction paradigms.

These observations are consistent with structural properties of the algorithm characterized in Supplementary S1 (summarized in Table S2). Two properties plausibly underlie the cross-domain transferability collapse. Intermediate Fourier magnitude normalization (Property 1) replaces the spectral magnitude profile of the phase-modulated field with a unit-valued spectrum at every iteration. The final magnitude operation discards phase (Property 3), so the released image does not retain the complex-valued final state. A third property bears on reconstruction-based recovery: per-image stochasticity from independent phase initializations and mask draws (Property 2) limits multi-realization averaging as a recovery strategy. Finally, the exponentially scaling latent mask configuration space (Property 4) is consistent with paired-supervision inversion failing to converge to a unique pre-image, even under direct training exposure to all evaluation subjects.

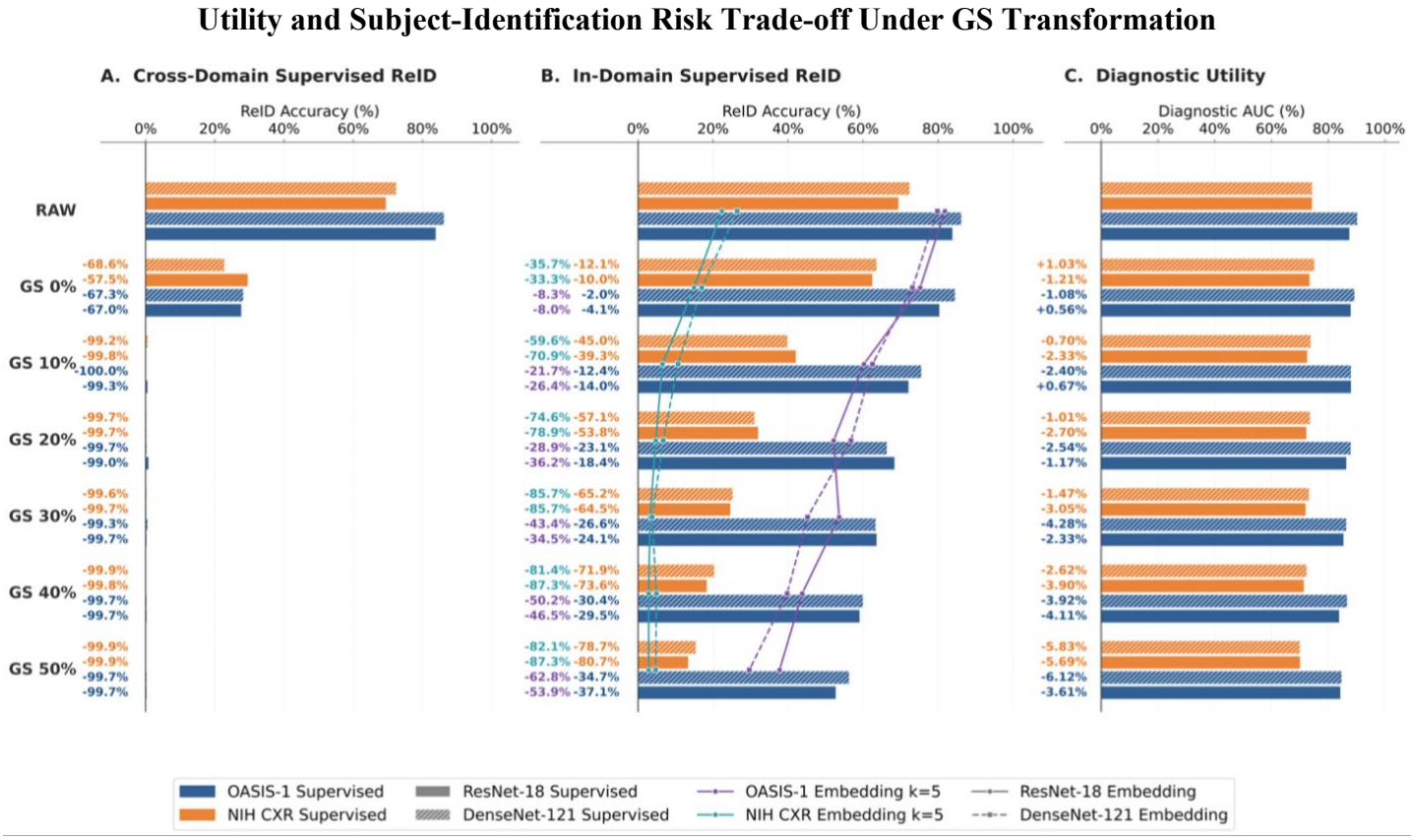


Figure 6 — Joint summary of diagnostic utility and subject-identification risk across GS masking levels. (A) Cross-domain supervised re-identification Top-1 accuracy, where classifiers are trained on GS-transformed images and evaluated on RAW images. (B) In-domain supervised re-identification Top-1 accuracy with embedding-based re-identification Rank-1 accuracy (k=5) overlay lines; embedding-based re-identification is not applicable under cross-domain conditions. (C) Diagnostic AUC under matched GS training and evaluation conditions. Percentage changes from RAW baseline are annotated per condition. Subject-identification risk declines substantially across all measurement conditions while diagnostic utility changes remain within 5.5% at GS-50%, reflecting a consistent empirical asymmetry across datasets and architectures.

Within the conditions evaluated, varying masking strength produces a range of empirical operating points between subject-identification risk reduction and diagnostic signal retention. GS operates as a training-free, input-level preprocessing step applicable independently of downstream models and without

modification to existing training pipelines. Whether the observed operating range is sufficient for any specific data-sharing scenario depends on that scenario's utility requirements, threat model, and acceptable adversary classes.

5. Limitations and Future Work

Findings are empirical and scoped to the conditions tested. They do not constitute formal privacy guarantees of the kind provided by frameworks such as differential privacy, and robustness against unevaluated conditions cannot be assumed. Comparative evaluation against alternatives such as defacing or differential privacy is identified as future work, building on a preliminary investigation [23].

The supervised re-identification designs are controlled conservative conditions and do not reflect typical data-sharing scenarios. Residual in-domain structure does not transfer to RAW images. However, the precise relationship between this residual structure and the original anatomical identity representation is not directly characterized here.

Several adversary classes remain unevaluated. Post-training evaluations such as model inversion and membership inference are scoped to follow-up work [23]. Foundation-model linkage [5] is a further unevaluated condition. Reduced re-identification under task-trained embeddings does not imply equivalent reduction under foundation-model embeddings, which differ in identity sensitivity.

6. Conclusion

This study characterizes empirical effects of GS transformation across two modalities and four measurement conditions. Diagnostic performance decreases modestly under increasing transformation strength, whereas subject-identification risk declines more sharply across all four adversary classes. Reconstruction-based analyses show that visually plausible images can be generated from GS-transformed inputs without recovering subject identity. This holds even when reconstruction models have direct training exposure to the RAW anatomical appearance of all evaluation subjects. Reconstruction quality alone is not a reliable indicator of subject-identification risk. GS operates as a training-free, input-level transformation compatible with standard training pipelines without architectural modification. Results reflect empirical characterization under assessed conditions and do not constitute formal privacy guarantees.

Supplementary Materials: The following supporting information can be downloaded at: PLACEHOLDER-FOR-EDITORIAL-OFFICE. Supplementary Material S1: Algorithm Specification and Structural Characterization of the GS Transformation; Supplementary Material S2: GS Parameter Sensitivity, Stochastic Stability, and Computational Profiling; Supplementary Material S3: Datasets, Cohort Construction, Preprocessing, and GS Dataset Generation; Supplementary Material S4: Diagnostic Utility Experiment Details; Supplementary Material S5: Supervised Re-Identification Experiment Details; Supplementary Material S6: Embedding-Based Re-Identification Details; Supplementary Material S7: Reconstruction-Based Inversion Details; Supplementary Material S8: Statistical Analysis and Confidence Intervals; Supplementary Material S9: Reproducibility Checklist and Output Mapping; Figure S1: Representative GS-transformed radiographs across masking levels; Figure S2: Fourier-domain visualization of the GS pipeline; Figure S3: Image-level stochasticity of GS transformation under repeated random realizations; Figure S4: Non-injectivity of the final magnitude operation; Figure S5: Mask configuration space and entropy characterization; Figure S6: GS iteration convergence analysis; Figure S7: GS stochastic stability under aggregated sample sizes; Figure S8: GS runtime profiling across image resolutions and batch sizes; Table S1: GS transformation notation; Table S2: Structural properties of the modified GS transformation; Table S3: Computational environment; Table S4: Eleven retained NIH chest X-ray classes with per-class and per-partition counts; Table S5: Cross-experiment split summary; Table S6:

Diagnostic utility training configuration; Table S7: Supervised re-identification training configuration; Table S8: Mapping from main-manuscript figures and tables to supplementary procedure sections and released-code modules.

Author Contributions: Removed for double-blind review.

Funding: Removed for double-blind review.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The experiments in this study use publicly available medical imaging datasets. Structural T1-weighted MRI scans from the OASIS-1 cohort are used to evaluate binary dementia classification and subject-level re-identification risk [12]. In addition, frontal chest radiographs from the NIH-CXR dataset are used for multi-label pathology classification and identity leakage evaluation [13]. All datasets were accessed under terms permitting academic research use and were processed using standard preprocessing procedures. All code required to reproduce the results is publicly available at: Anonymized GitHub Repository.
The diffusion posterior sampling adversary uses the pretrained model and inference framework from Chung et al. [18], available at https://github.com/DPS2022/diffusion-posterior-sampling. The pretrained checkpoint (ffhq_10m.pt) was obtained directly from that repository and used without modification as the base model prior to domain-specific fine-tuning. Default inference configurations follow the repository specifications.

Acknowledgments: Removed for double-blind review.

Conflicts of Interest: The authors declare no competing interests.

Abbreviations

The following abbreviations are used in this manuscript:

Order	Abbreviation	Expansion	First appears in
1	GS	Gerchberg-Saxton	Main, Title / Abstract
2	MRI	Magnetic Resonance Imaging	Main, Abstract
3	AUC	Area Under the (ROC) Curve	Main, Abstract
4	CI	Confidence Interval	Main, Abstract
5	SSIM	Structural Similarity Index Measure	Main, Abstract
6	FE	Feature Embedding	Main, Table 2
7	PSNR	Peak Signal-to-Noise Ratio	Main, Section 2.5
8	DPS	Diffusion Posterior Sampling	Main, Section 3.3
9	FFHQ	Flickr-Faces-HQ	Main, Section 3.3
10	DFT	Discrete Fourier Transform	Supplement S1.2
11	LPIPS	Learned Perceptual Image Patch Similarity	Supplement S2.2
12	CV	Cross-Validation	Supplement S3
13	MPR	Motion-Corrected Reconstruction	Supplement S3
14	GPU	Graphics Processing Unit	Supplement S2
15	CPU	Central Processing Unit	Supplement S2
16	RAM	Random-Access Memory	Supplement S2
17	OOM	Out-Of-Memory	Supplement S2
18	CUDA	Compute Unified Device Architecture	Supplement S4
19	AMP	Automatic Mixed Precision	Supplement S4

Order	Abbreviation	Expansion	First appears in
20	RGB	Red, Green, Blue	Supplement S4
21	AdamW	Adam with decoupled Weight decay	Supplement S4
22	RNG	Random Number Generator	Supplement S4
23	ReLU	Rectified Linear Unit	Supplement S6
24	TAR	True Accept Rate	Supplement S6
25	FAR	False Accept Rate	Supplement S6

References

[1] A. Esteva *et al.*, "A guide to deep learning in healthcare," (in eng), *Nat Med,* vol. 25, no. 1, pp. 24-29, Jan 2019, doi: 10.1038/s41591-018-0316-z.

[2] S. M. McKinney *et al.*, "International evaluation of an AI system for breast cancer screening," *Nature,* vol. 577, no. 7788, pp. 89-94, 2020/01/01 2020, doi: 10.1038/s41586-019-1799-6.

[3] E. E. L. Buimer *et al.*, "De-identification procedures for magnetic resonance images and the impact on structural brain measures at different ages," (in eng), *Hum Brain Mapp,* vol. 42, no. 11, pp. 3643-3655, Aug 1 2021, doi: 10.1002/hbm.25459.

[4] C. G. Schwarz *et al.*, "Identification of Anonymous MRI Research Participants with Face-Recognition Software," (in eng), *N Engl J Med,* vol. 381, no. 17, pp. 1684-1686, Oct 24 2019, doi: 10.1056/NEJMc1908881.

[5] G. Nebbia *et al.*, "Re-identification of patients from imaging features extracted by foundation models," (in eng), *NPJ Digit Med,* vol. 8, no. 1, p. 469, Jul 22 2025, doi: 10.1038/s41746-025-01801-0.

[6] M. Abadi *et al.*, "Deep Learning with Differential Privacy," presented at the Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 2016. [Online]. Available: https://doi.org/10.1145/2976749.2978318.

[7] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends in Theoretical Computer Science,* vol. 9, 01/01 2013, doi: 10.1561/0400000042.

[8] A. Ziller, D. Usynin, R. Braren, M. Makowski, D. Rueckert, and G. Kaissis, "Medical imaging deep learning with differential privacy," (in eng), *Sci Rep,* vol. 11, no. 1, p. 13524, Jun 29 2021, doi: 10.1038/s41598-021-93030-0.

[9] A. Chartsias *et al.*, "Disentangled representation learning in cardiac image analysis," (in eng), *Med Image Anal,* vol. 58, p. 101535, Dec 2019, doi: 10.1016/j.media.2019.101535.

[10] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proceedings of the IEEE,* vol. 69, pp. 529-541, 1980.

[11] R. W. Gerchberg, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik,* vol. 35, pp. 237-246, 1972.

[12] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults," *Journal of Cognitive Neuroscience,* vol. 19, no. 9, pp. 1498-1507, 2007, doi: 10.1162/jocn.2007.19.9.1498.

[13] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3462-3471, 2017.

[14] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," (in eng), *Med Image Anal,* vol. 42, pp. 60-88, Dec 2017, doi: 10.1016/j.media.2017.07.005.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[16] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2261-2269, 2016.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Cham, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., 2015// 2015: Springer International Publishing, pp. 234-241.

[18] H. Chung, J. Kim, M. T. McCann, M. L. Klasky, and J. C. Ye, "Diffusion Posterior Sampling for General Noisy Inverse Problems," *ArXiv,* vol. abs/2209.14687, 2022.

[19] A. Kazerouni *et al.*, "Diffusion models in medical imaging: A comprehensive survey," (in eng), *Med Image Anal,* vol. 88, p. 102846, Aug 2023, doi: 10.1016/j.media.2023.102846.

[20] G. Kaissis *et al.*, "End-to-end privacy preserving deep learning on multi-institutional medical imaging," *Nature Machine Intelligence,* vol. 3, pp. 473-484, 05/24 2021, doi: 10.1038/s42256-021-00337-8.

[21] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb, "Solving inverse problems using data-driven models," *Acta Numerica,* vol. 28, pp. 1-174, 2019, doi: 10.1017/S0962492919000059.

[22] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep Convolutional Neural Network for Inverse Problems in Imaging," *Trans. Img. Proc.,* vol. 26, no. 9, pp. 4509–4522, 2017, doi: 10.1109/tip.2017.2713099.

[23] Anonymous citation withheld for double-blind review.

Supplementary Materials

Supplementary Material S1: Algorithm Specification and Structural Characterization of the GS Transformation

S1.1 Purpose and Scope

This supplementary section provides the algorithmic specification and structural characterization of the modified Gerchberg-Saxton transformation used in this study. The goal of this section is to define the transformation precisely, clarify the notation used in the algorithm, and identify structural properties that follow directly from the procedure itself. Subsequent supplementary sections provide the experimental implementation details, parameter settings, dataset-specific processing, iteration-convergence analysis (S2.2), stochastic-stability analysis (S2.3), and step-by-step reproducibility information.

The structural properties described here should not be interpreted as formal privacy guarantees. In particular, the transformation is not claimed to satisfy (ε, δ)-differential privacy, and no parameter is calibrated to the global sensitivity of a query function with respect to individual subjects. The masking probability p and iteration count I are algorithmic strength parameters that control the degree of spectral perturbation and repeated phase updating applied to each image. Their empirical effects are evaluated in the main manuscript through subject-identification risk, reconstruction-based inversion, embedding-based linkage, and diagnostic utility experiments.

The purpose of this section is therefore to distinguish what follows directly from the algorithmic construction from what is evaluated empirically in the main study and in the additional supplementary sensitivity analyses.

Table S1. GS Transformation Notation

Notation	Meaning	Notation	Meaning
$\boldsymbol{X}$	Input image	ϕ_t	Spatial phase field at iteration t
$\boldsymbol{H}, \boldsymbol{W}$	Image height, width	$\Phi_t = \exp(i\phi_t)$	Complex unit-magnitude phase field at iteration t
$\boldsymbol{N} = \boldsymbol{H} \cdot \boldsymbol{W}$	Total number of image pixels	Z_t	Phase-modulated image-domain field at iteration t
$\boldsymbol{I}$	Number of GS iterations	G_t	Fourier-domain representation of Z_t
$\boldsymbol{p}$	Per-location masking probability	$\tilde{G}_t$	Magnitude-normalized Fourier-domain field
$\boldsymbol{F_H}, \boldsymbol{F_W}$	Unitary DFT matrices along the image height and width dimensions	M_t	Random binary frequency mask at iteration t
$\boldsymbol{F_H^{-1}}, \boldsymbol{F_W^{-1}}$	Inverse DFT matrices	$\hat{G}_t$	Masked normalized Fourier-domain field
$\odot$	Element-wise multiplication	$\hat{Z}_t$	Inverse DFT of $\hat{G}_t$
$\oslash$	Element-wise division	$Y = \lvert\hat{Z}_I\rvert$	final GS-transformed output image
$\lvert\cdot\rvert$	Complex modulus or magnitude	ϵ	Numerical tolerance used to avoid division by zero
$\angle(\cdot)$	Complex phase angle	$\mathcal{N}_\epsilon$	Magnitude-normalization operator with numerical tolerance

S1.2 Notation

Let $X \in \mathbb{R}^{H \times W}$ denote a single input image with height H, width W, and $N = H \cdot W$ total pixels. Let $I \in \mathbb{N}$ denote the number of GS iterations and $p \in [0,1)$ denote the per-location masking probability. Let $F_H \in \mathbb{C}^{H \times H}$ and $F_W \in \mathbb{C}^{W \times W}$ denote unitary discrete Fourier transform matrices scaled by $1/\sqrt{H}$ and $1/\sqrt{W}$, respectively, with F_H^{-1} and F_W^{-1} denoting their inverses. Because the DFT matrix is symmetric ($F = F^T$), the standard 2D-DFT expression $F_H \cdot X \cdot F_W^T$ is equivalent to $F_H \cdot \mathrm{X} \cdot F_W$, which is the form used in Algorithm 1 (line 8). The same symmetry applies to the inverse pair on line 12. Element-wise multiplication and division are denoted by $\odot$ and $\oslash$, respectively. The complex modulus and phase are denoted by $|\cdot|$ and $\angle(\cdot)$. The distributions $\mathrm{Uniform}(a, b)$ and $\mathrm{Bernoulli}(q)$ denote the continuous uniform and Bernoulli distributions, respectively. The main notation used throughout Algorithm 1 and the structural

characterization is summarized in Table S1. For numerical completeness, define the magnitude-normalization operator

$$\mathcal{N}_\epsilon(G)[u,v] = \begin{cases} G[u,v] \,/\, |\, G[u,v]\,|, & |\, G[u,v]\,| > \epsilon, \\ 0, & |\, G[u,v]\,| \leq \epsilon, \end{cases}$$

where $\epsilon \geq 0$ is a small numerical tolerance. When $|\, G[u,v]\,| > \epsilon$, this is equivalent to the element-wise operation $G \oslash |\, G\,|$. In exact mathematical notation, $\epsilon = 0$ is sufficient whenever $G[u,v] \neq 0$. In numerical implementations, a small positive value avoids division by zero or unstable division by near-zero coefficients. Similarly, the phase of zero-valued entries is assigned by convention, for example $\angle(0) = 0$. Hat notation is used as follows: $\tilde{G}_t$ (with tilde) denotes the magnitude-normalized field produced at line 9 of Algorithm 1, $\hat{G}_t$ (with hat) denotes the masked field produced at line 11 after applying M_t to $\tilde{G}_t$ (with tilde), and $\hat{Z}_t$ (with hat) denotes the inverse DFT of $\hat{G}_t$ (with hat) produced at line 12; the final output $Y = |\, \hat{Z}_I\,|$ (with hat) refers to the magnitude of this masked-then-inverted quantity, not of the normalized-only quantity.

Algorithm 1. Modified Gerchberg-Saxton Image Transformation

1:	function GS-TRANSFORM(X, I, p, ε)
2:	$\boldsymbol{F_H} \leftarrow$ DFT matrix of order $\boldsymbol{H}$, scaled by $\boldsymbol{1/\sqrt{H}}$
3:	$\boldsymbol{F_W} \leftarrow$ DFT matrix of order $\boldsymbol{W}$, scaled by $\boldsymbol{1/\sqrt{W}}$
4:	$\boldsymbol{\varphi_0[h,w]} \sim \boldsymbol{Uniform(0, 2\pi)}$ for all $\boldsymbol{h} \in \{1,\ldots,H\}, \boldsymbol{w} \in \{1,\ldots,W\}$
5:	$\boldsymbol{\Phi_0} \leftarrow \boldsymbol{e^{i\varphi_0}}$
6:	for $\boldsymbol{t} \leftarrow \boldsymbol{1}$ to $\boldsymbol{I}$ do
7:	$\boldsymbol{Z_t} \leftarrow \boldsymbol{X} \odot \boldsymbol{\Phi_{t-1}}$
8:	$\boldsymbol{G_t} \leftarrow \boldsymbol{F_H} \cdot \boldsymbol{Z_t} \cdot \boldsymbol{F_W}$
9:	$\mathbf{G_t} \leftarrow \boldsymbol{N_\varepsilon}(\mathbf{G_t})$ # equivalent to $\mathbf{G_t} \oslash \lvert\mathbf{G_t}\rvert$ where $\lvert\mathbf{G_t}\rvert > \boldsymbol{\varepsilon}$
10:	$\mathbf{M_t[h,w]} \sim \boldsymbol{Bernoulli(1-p)}$ i.i.d. for all h, w
11:	$\hat{\mathrm{G}}_t \leftarrow \mathbf{G_t} \odot \mathbf{M_t}$
12:	$\boldsymbol{Z_t} \leftarrow \boldsymbol{F_H^{-1}} \cdot \hat{\mathrm{G}}_t \cdot \boldsymbol{F_w^{-1}}$
13:	$\boldsymbol{\varphi_t} \leftarrow \angle(\hat{\mathrm{Z}}_t)$
14:	$\boldsymbol{\Phi_t} \leftarrow \boldsymbol{e^{i\varphi_t}}$
15:	end for
16:	return $\lvert\hat{\mathrm{Z}}_I\rvert$
17:	end function

For audiences less familiar with the domain, this notation separates the major stages of the GS transformation. Z_t is the image after phase modulation, G_t is its Fourier-domain representation, $\tilde{G}_t$ (with tilde) is the magnitude-normalized Fourier representation, $\hat{G}_t$ (with hat) is the same representation after stochastic masking, and $\hat{Z}_t$ (with hat) is the masked representation transformed back into the image domain. The released output Y is the magnitude of the final $\hat{Z}_t$ (with hat), which is why the output should be interpreted as the magnitude of the masked-then-inverted representation rather than the magnitude of the normalized-only Fourier field.

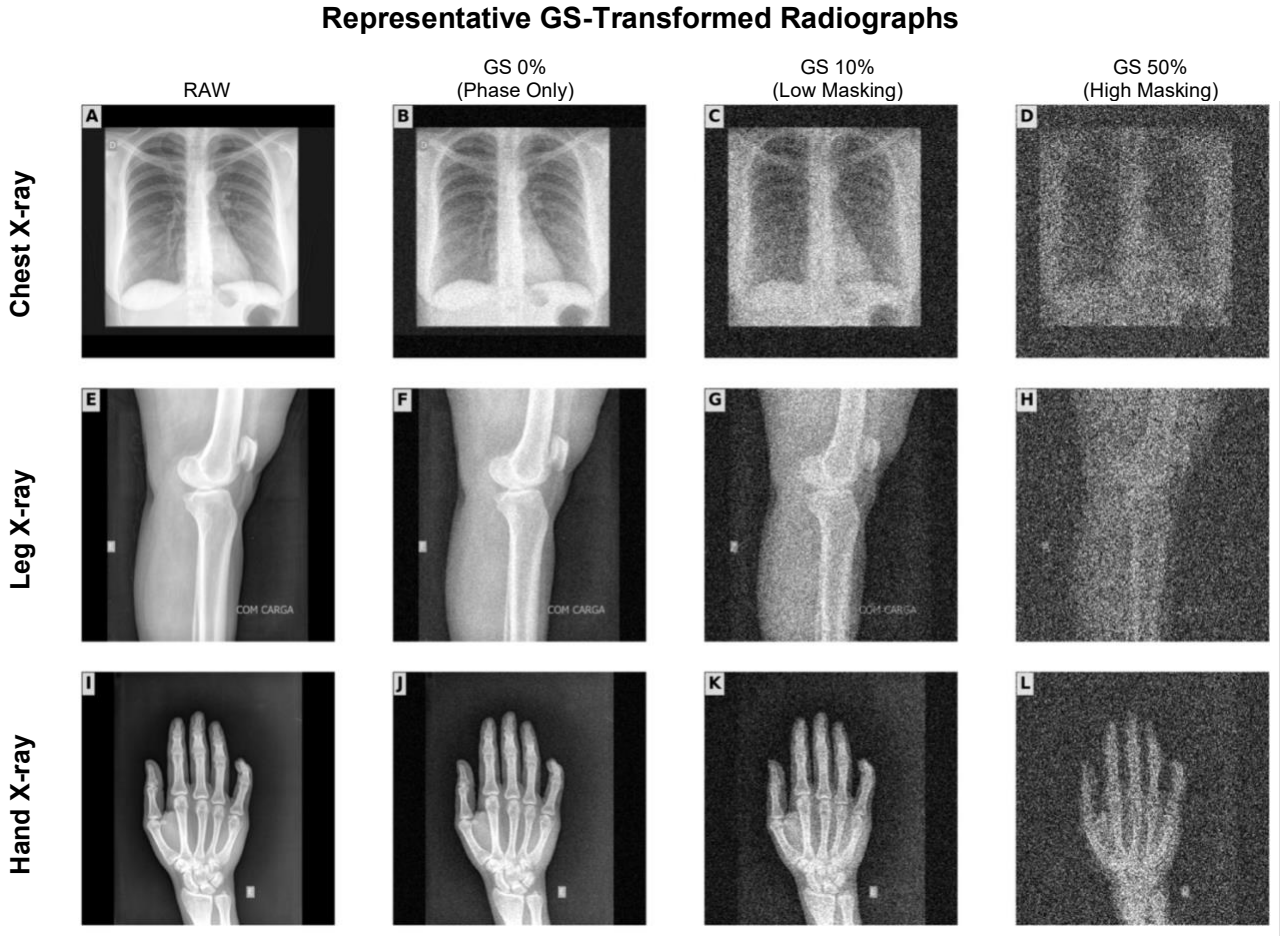
S1.3 Algorithm

In line 9, $\mathcal{N}_\epsilon(G_t)$ denotes magnitude normalization with numerical tolerance, equivalent to

$$G_t \oslash |\, G_t\,| \text{ wherever } |\, G_t\,| > \epsilon.$$

The transformation is applied independently to each image. It does not use subject labels, class labels, anatomical priors, segmentation masks, or model-specific information. Two sources of stochasticity are

present: the initial phase field on line 4, drawn independently for each image and run, and the binary frequency mask on line 10, drawn independently at each iteration. The output on line 16 discards the phase of the final complex-valued image-domain field and retains only its magnitude. In the experiments reported in the main manuscript, transformed images are generated once per experimental condition and reused across downstream evaluations, so that observed differences reflect the chosen transformation condition rather than repeated stochastic resampling.



Supplementary Figure S1. Chest, leg, and hand radiographs are shown before transformation and after GS transformation at $p = 0$, $p = 0.1$, and $p = 0.5$ using 50 iterations ($I = 50$). GS-0% reflects phase perturbation without frequency masking, whereas increasing p introduces progressively stronger stochastic removal of normalized frequency components. These examples illustrate the qualitative image-domain effect of Algorithm 1 across increasing transformation strength.

Representative outputs generated by Algorithm 1 are shown in Supplementary Figure S1. The examples illustrate how the same transformation procedure produces visibly different image-domain representations as masking strength increases from GS-0% to GS-50%. GS-0% reflects phase perturbation without frequency masking, whereas GS-10% and GS-50% introduce progressively stronger stochastic removal of normalized frequency components.

Operationally, Algorithm 1 describes GS as a randomized preprocessing operation applied before model training or evaluation. The image is repeatedly phase-modulated, transformed into the Fourier domain, magnitude-normalized, partially masked, returned to the image domain, and updated through its phase. The released image is not the intermediate Fourier representation itself, but the magnitude of the final masked-and-inverted complex image-domain field.

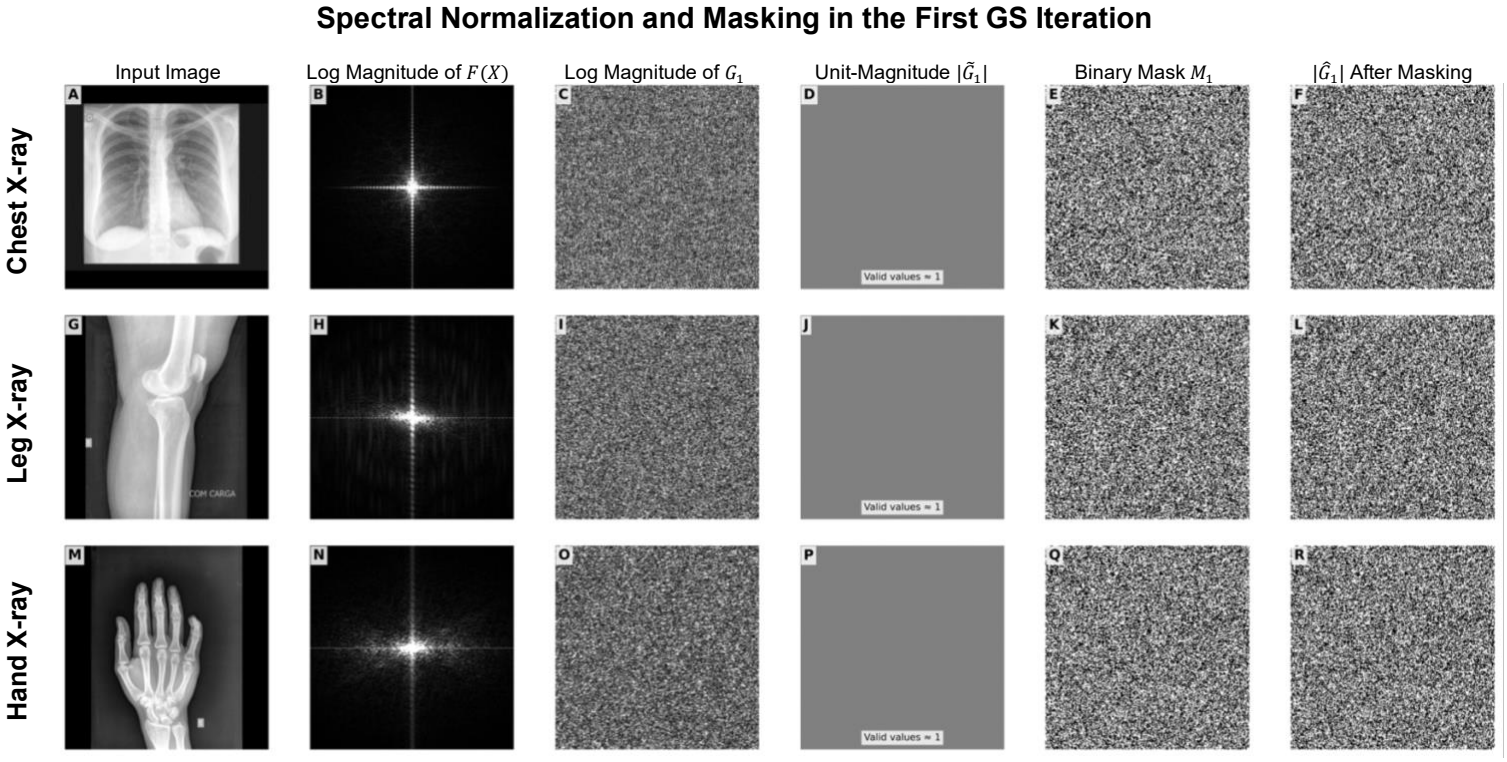
S1.4 Structural Property 1: Intermediate Fourier Magnitude Normalization

At each iteration, line 9 of Algorithm 1 replaces the Fourier magnitude of the current phase-modulated field $Z_t = X \odot \Phi_{t-1}$ with a unit-valued spectrum at all coefficients whose magnitude exceeds the numerical tolerance. Specifically, for every coefficient (u, v) satisfying $| G_t[u, v] | > \epsilon$,

$$| \tilde{G}_t[u, v] | = 1.$$

For coefficients satisfying $| G_t[u, v] | \leq \epsilon$, the normalization operator assigns $\tilde{G}_t[u, v] = 0$ by convention. This property follows directly from the definition of $\mathcal{N}_\epsilon$. If $| G_t[u, v] | > \epsilon$, then

$$| \tilde{G}_t[u, v] | = \left| \frac{G_t[u, v]}{| G_t[u, v] |} \right| = 1$$



Supplementary Figure S2. For each representative radiograph, the input image is shown alongside the log Fourier magnitude of the original image, the log magnitude of the first phase-modulated Fourier field G_1, the magnitude-normalized field $| \tilde{G}_1 |$, the sampled binary frequency mask M_1, and the masked normalized field $| \hat{G}_1 |$. The $| \tilde{G}_1 |$ panels are expected to appear spatially uniform because line 9 of Algorithm 1 normalizes valid Fourier coefficients to approximately unit magnitude. The final column shows the effect of masking after normalization, demonstrating that retained and removed frequency locations are determined by M_1 before inverse transformation.

Thus, the algorithm explicitly removes the magnitude profile of the intermediate Fourier representation before masking and inverse transformation. Importantly, this statement applies to the Fourier magnitude of the intermediate phase-modulated representation Z_t, not to all information contained in the original image X. The input image continues to influence the phase of G_t, the inverse-transformed field $\hat{Z}_t$, and the subsequent phase update. Therefore, the correct interpretation is that the algorithm does not pass the raw Fourier magnitude of Z_t forward at the normalization step. It should not be interpreted as proof that all information derived from X has been removed.

Supplementary Figure S2 illustrates this normalization step using representative radiographs. The original Fourier magnitude and the phase-modulated Fourier magnitude retain visible spectral structure, whereas

after magnitude normalization $|\tilde{G}_1|$ becomes approximately uniform. The subsequent mask M_1 then determines which normalized frequency locations are retained or removed before the inverse transform. This figure visually separates the normalization operation from the masking operation, which is important because the two steps play distinct roles in Algorithm 1.

The key interpretation is that line 9 does not erase the entire image or prove that all identity-related information is removed. Rather, it prevents the magnitude profile of the current Fourier representation from being directly carried forward. The image still influences the transformation through phase and through subsequent updates, but the raw Fourier magnitude of the intermediate field is replaced by a unit-magnitude representation before masking and inversion.

S1.5 Structural Property 2: The Transformation Is Stochastic for a Fixed Input

Following Algorithm 1, for a fixed input image X, the transformed output can be written as

$$Y = f(X, \phi_0, M_1, \ldots, M_I),$$

where ϕ_0 is the initial random phase field and $M_1, \ldots, M_I$ are independently sampled binary frequency masks. Therefore, unless the random seed and all mask realizations are fixed, the output is a random variable even when the input image is held constant. The number of independently sampled random quantities per image is

$$N + I \cdot N = N(1 + I),$$

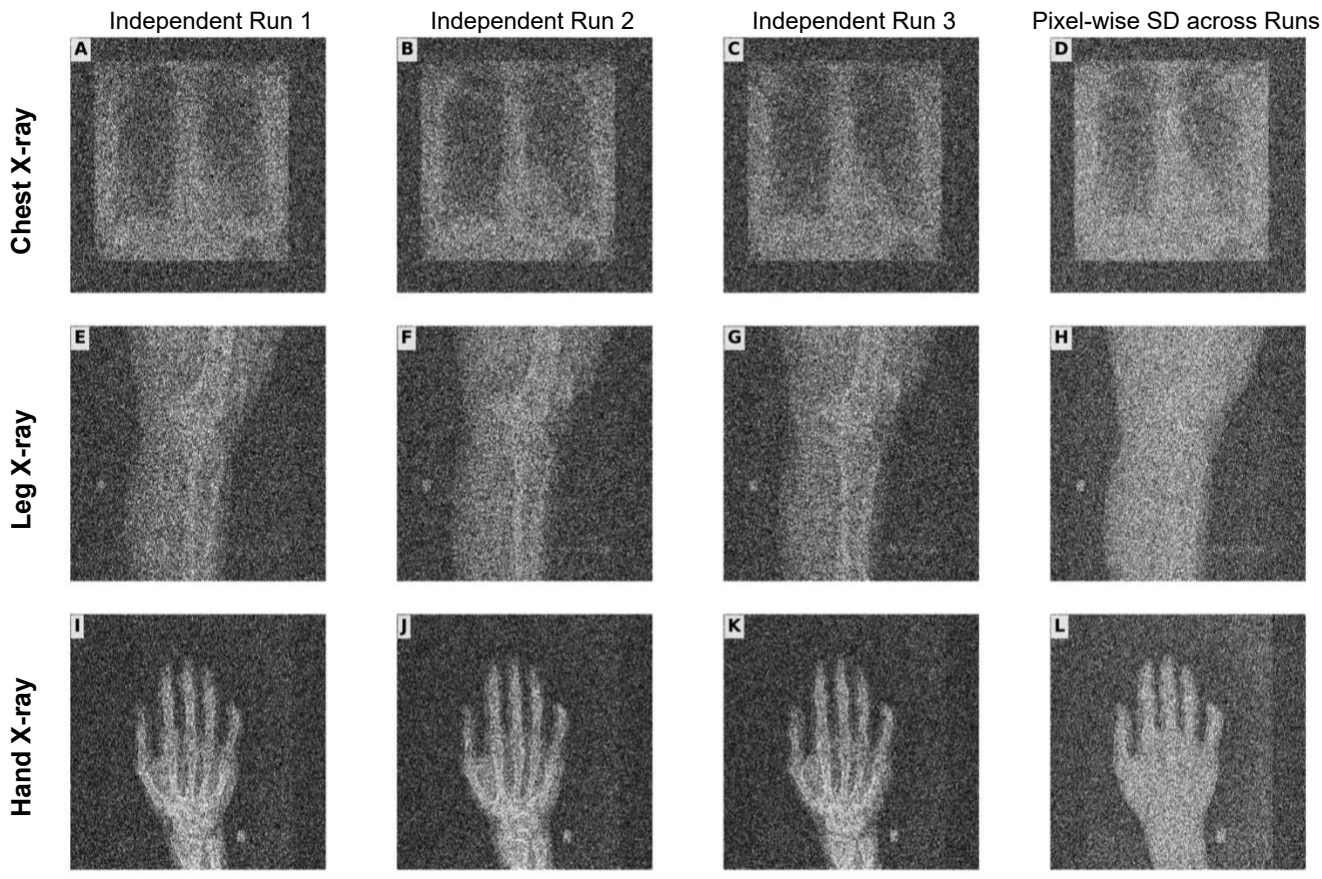
where $N = H \cdot W$. The first term corresponds to the continuous initial phase values, and the second term corresponds to binary mask entries sampled across all iterations. For the binary masks alone, the entropy of the latent mask process is $H(M_{1:I}) = I \cdot H \cdot W \cdot H_b(p)$, where

$$H_b(p) = -p \cdot \log_2(p) - (1 - p) \cdot \log_2(1 - p)$$

is the binary entropy function, measured in bits because the logarithm is base 2. At the boundary values $p = 0$ and $p = 1$, entropy terms of the form $p \cdot \log_2(p)$ are assigned their limiting value of zero. This convention reflects the fact that a deterministic mask has zero entropy: when $p = 0$, no coefficients are masked, and when $p = 1$, all coefficients are masked. This expression describes the entropy of the unobserved mask process, not necessarily the entropy of the final image Y. Because Y is a deterministic function of X, ϕ_0, and $M_{1:I}$, the entropy of the output cannot be assumed equal to the entropy of the latent random variables. Some random choices may produce similar or even identical outputs. Therefore, the conservative claim is that the algorithm introduces latent stochasticity that scales with image size and iteration count. Empirical variability of the output can be estimated by repeatedly transforming the same fixed input under independent random seeds.

Supplementary Figure S3 illustrates this fixed-input stochasticity. The same radiograph is transformed multiple times using the same GS parameters but independent random draws. The resulting images differ at the pixel level, and the pixelwise standard deviation map highlights where this run-to-run variability appears. This supports the interpretation that GS is stochastic at the individual-image level, even though the main experiments generate each transformed dataset once per condition and reuse it consistently across evaluations; the stochastic-stability analysis in S2.3 evaluates whether aggregate behavior is stable across independent random seeds.

Conceptually, this means that GS does not define a single deterministic output for a given input image unless the random phase initialization and all masks are fixed. The important claim here is not that the final transformed image has a specific guaranteed entropy, but that the algorithm contains independent latent randomness whose dimensionality grows with image size and iteration count. This stochasticity is part of the transformation mechanism, while its empirical consequences for re-identification and utility are evaluated in the main experiments and in the stochastic-stability analysis in S2.3.



Supplementary Figure S3. The same input radiograph is transformed repeatedly using identical GS parameters but independent random draws of the initial phase field and per-iteration masks. The first three columns show representative independent outputs, and the final column shows the pixelwise standard deviation across repeated runs. The figure illustrates that GS is stochastic at the image level: repeated transformations of the same image can produce different pixel-level outputs unless the random initialization and masks are fixed. This variability motivates the stochastic-stability analysis reported in later supplementary sections, where aggregate behavior is assessed across repeated random seeds.

S1.6 Structural Property 3: The Final Magnitude Operation Discards Phase

The final output is $Y = |\hat{Z}_I|$. This operation discards the phase of the final complex-valued field $\hat{Z}_I$. For any phase field $\theta \in [0,2\pi)^{H\times W}$,

$$|\hat{Z}_I \odot \exp(i\theta)| = |\hat{Z}_I|.$$

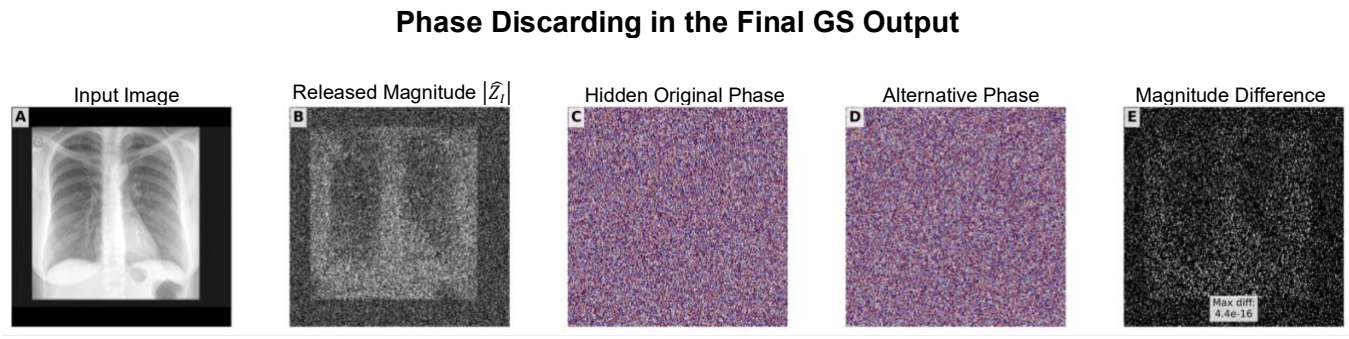
Thus, many distinct complex-valued final states map to the same real-valued output image. This establishes non-injectivity of the final magnitude operation with respect to the complex final state.

This statement should be interpreted carefully. It does not prove that the original input image X is unrecoverable under all adversarial models, nor does it provide a formal lower bound on reconstruction error. It only establishes that phase information present in the final complex representation is not included in the released transformed image.

Supplementary Figure S4 illustrates this phase-discarding property. The released magnitude image remains unchanged when the hidden phase of the final complex-valued field is replaced with an alternative phase field, up to numerical precision. This visual demonstration supports the statement that the released output

Y does not contain the full complex final state, even though it does not by itself establish reconstruction impossibility.

This property should therefore be understood as a statement about information discarded at the final representation step. The released GS image contains the magnitude of $\hat{Z}_I$, but not its phase. Consequently, different complex-valued final states can correspond to the same released magnitude image. The empirical reconstruction experiments in the main manuscript then evaluate whether adversarial reconstruction methods can recover identity-bearing structure despite this phase-discarding operation.



Supplementary Figure S4. The GS transformation produces a final complex-valued image-domain field $\hat{Z}_I$, but the released output is only its magnitude, $Y = |\hat{Z}_I|$. Replacing the hidden phase with an alternative phase field changes the underlying complex-valued field while preserving the same magnitude image up to numerical precision. This demonstrates that the final released GS image does not retain the phase of the final complex representation. The figure illustrates non-injectivity of the final magnitude operation with respect to the complex final state but does not by itself establish that the original input image is unrecoverable under all adversarial models.

S1.7 Structural Property 4: The Latent Mask Configuration Space Scales Exponentially

At each iteration, one binary mask value is sampled independently for each frequency location. Across I iterations and $H \cdot W$ frequency locations, the number of possible binary mask histories is $2^{I \cdot H \cdot W}$. The expected number of masked frequency-location slots is

$$\mathbb{E}[\sum_{t=1}^{I} \sum_{u=1}^{H} \sum_{v=1}^{W} \mathbf{1}\{M_t[u,v] = 0\}] = p \cdot I \cdot H \cdot W$$

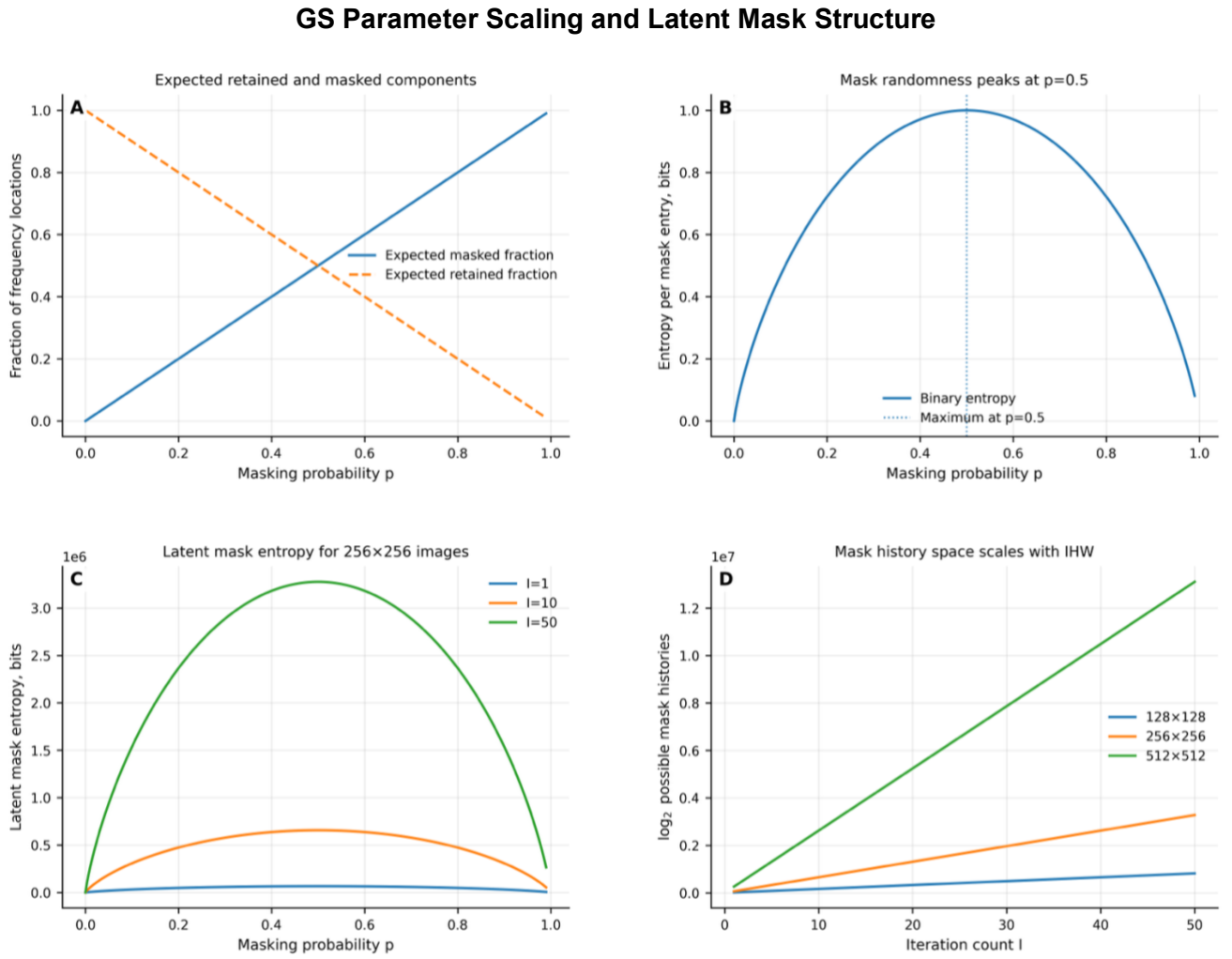
where $1\{\cdot\}$ denotes the indicator function, defined as

$$\mathbf{1}\{M_t[u,v] = 0\} = \begin{cases} 1, & \text{if } M_t[u,v] = 0 \\ 0, & \text{if } M_t[u,v] \neq 0 \end{cases}$$

This describes the size of the latent random mask space and the expected number of removed normalized frequency components. It should not be interpreted as a proof that 2^{pIHW} different mask histories are consistent with every observed output. The final output is affected by the mask history through the inverse transform and subsequent phase updates, so some information about the mask realization may be reflected indirectly in the output.

The conservative interpretation is that the adversary does not observe a mask history drawn from an exponentially large latent configuration space. This contributes to the stochastic underdetermination of the released image, but it is not equivalent to a formal pre-image lower bound.
Supplementary Figure S5 visualizes this scaling behavior. The figure shows that the expected masked fraction increases linearly with p, while the entropy of the binary mask process is maximized at $p = 0.5$. It also shows that latent mask entropy increases with iteration count and that the total number of possible

mask histories scales with $I \cdot H \cdot W$ on the $\log_2$ scale. These plots support the interpretation of p and I as transformation-strength parameters rather than formal privacy budgets.



Supplementary Figure S5. Panel A shows that increasing the masking probability p linearly increases the expected fraction of masked frequency locations while decreasing the expected retained fraction. Panel B shows that binary mask entropy is maximized at $p = 0.5$ and decreases as the mask becomes deterministic near $p = 0$ or $p = 1$. Panel C shows that latent mask entropy increases with iteration count for a fixed image size. Panel D shows that the number of possible binary mask histories grows with $I \cdot H \cdot W$ on the $\log_2$ scale. Together, these plots support interpreting p and I as algorithmic transformation-strength parameters rather than formal privacy budgets.

Practically, this means that GS introduces a large unobserved mask space, but this should not be confused with a formal guarantee of irreversibility. The algorithm samples from many possible mask histories, and the adversary does not observe the specific realization. However, the released output may still contain indirect effects of the sampled mask. Therefore, the scientifically justified claim is that the latent mask configuration space is large and grows rapidly with image size and iteration count, not that every output has a proven exponentially large pre-image set.

S1.8 Relationship Between Algorithmic Parameters and Empirical Evaluation

The masking probability p and iteration count I control distinct aspects of the transformation. The parameter p controls the expected fraction of normalized frequency components removed at each iteration. The

iteration count I controls how many rounds of spectral normalization, masking, inverse transformation, and phase updating are applied.

These parameters are privacy-relevant in an empirical sense because they control the strength of the perturbation applied before downstream modeling or adversarial evaluation. However, they are not privacy budgets. Increasing p increases the expected number of masked components, but the entropy of the binary mask process is maximized at $p = 0.5$, not at $p = 1$. As p approaches 1, the mask becomes increasingly deterministic because almost all coefficients are removed. Similarly, increasing I increases the number of repeated transformation steps, but this does not by itself establish a formal monotonic privacy guarantee.

Supplementary Figure S5 summarizes these parameter relationships. The expected masked fraction increases linearly with p, whereas binary mask entropy is non-monotonic and reaches its maximum at $p = 0.5$ *(S5-B)*. The same figure also shows how latent mask entropy increases with iteration count (*S5-C)* and how the possible mask-history space scales with $I \cdot H \cdot W$ *(S5-D)*. These observations support interpreting p and I as algorithmic transformation-strength parameters rather than formal privacy budgets.

In the main experiments, I is fixed at 50 and p is varied across the evaluated GS masking levels. This choice of I=50 is empirical and pragmatic rather than theoretically optimal. The iteration-convergence analysis in S2.2 shows that GS outputs stabilize by approximately this range, and the stochastic-stability analysis in S2.3 shows that aggregate behavior remains stable across random seeds. Accordingly, the main manuscript evaluates these parameters empirically rather than treating them as formal privacy controls. The relevant empirical questions are whether increasing transformation strength reduces re-identification, whether diagnostic utility remains stable, and whether reconstruction methods can recover identity-relevant information from transformed images.

For purposes of this study, p and I should therefore be understood as experimentally chosen transformation parameters. The value $I = 50$ is used because the sensitivity analyses in S2.2 and S2.3 support it as a stable and computationally practical operating point, not because it is a theoretically optimal stopping rule. The masking probability p is then varied to characterize how increasing transformation strength affects subject-identification risk, reconstruction behavior, and diagnostic utility under the evaluated conditions.

S1.9 Recommended Interpretation

The modified GS transformation should be interpreted as a randomized, image-domain preprocessing mechanism that repeatedly normalizes intermediate Fourier magnitudes, removes a randomly selected subset of normalized frequency components, updates spatial phase estimates, and releases only the magnitude of the final inverse-transformed representation.

These properties support the empirical motivation for evaluating GS as a re-identification mitigation transformation. They do not prove privacy, prevent all possible reconstruction attacks, or guarantee that identity information is absent. For that reason, the main claims of the manuscript should remain empirical: GS reduces measured re-identification under the evaluated threat models while preserving diagnostic utility within the tested conditions.

The four structural properties established in S1.4-S1.7 are summarized in Table S2, with each property tied to the figure (S1-S5) introduced alongside it.

The key interpretation is that these are structural properties of the transformation, not formal privacy guarantees. They explain why GS is scientifically reasonable to evaluate as a de-identification-oriented preprocessing mechanism, but the privacy-relevant conclusions of the manuscript come from the empirical re-identification, reconstruction, and diagnostic utility experiments rather than from these structural properties alone.

Table S2. Structural Properties of the Modified GS Transformation

Property	What is established	Scope
Intermediate Fourier magnitude normalization	The Fourier magnitude of the current phase-modulated field is replaced by a unit-valued spectrum at coefficients whose magnitude exceeds the numerical tolerance.	Applies to the intermediate representation Z_t, not to all information contained in X.
Fixed-input stochasticity	For a fixed input image, the output depends on the initial random phase field and independently sampled binary masks unless the random seed and masks are fixed.	Describes latent stochasticity and output variability, does not guarantee formal entropy.
Final magnitude phase discarding	The final output $Y = \lvert\hat{Z}_I\rvert$ discards the phase of the final complex-valued field.	Establishes non-injectivity of the final magnitude operation with respect to the complex final state.
Exponential latent mask configuration space	Across I iterations and $H \cdot W$ frequency locations, the number of possible binary mask histories is $2^{I \cdot H \cdot W}$.	Describes the latent mask space; does not provide a formal pre-image lower bound for every output observed.

Material S2: GS Parameter Sensitivity, Stochastic Stability, and Computational Profiling

S2.1 Overview

Supplementary Material S1 defined the modified Gerchberg-Saxton transformation and described its structural properties. This section evaluates implementation-level behavior that determines how the transformation is used in the main experiments. Specifically, we examine the convergence behavior of the iterative GS procedure, the stability of GS outputs under repeated stochastic realizations, and the computational cost of applying GS across image resolutions and batch sizes.

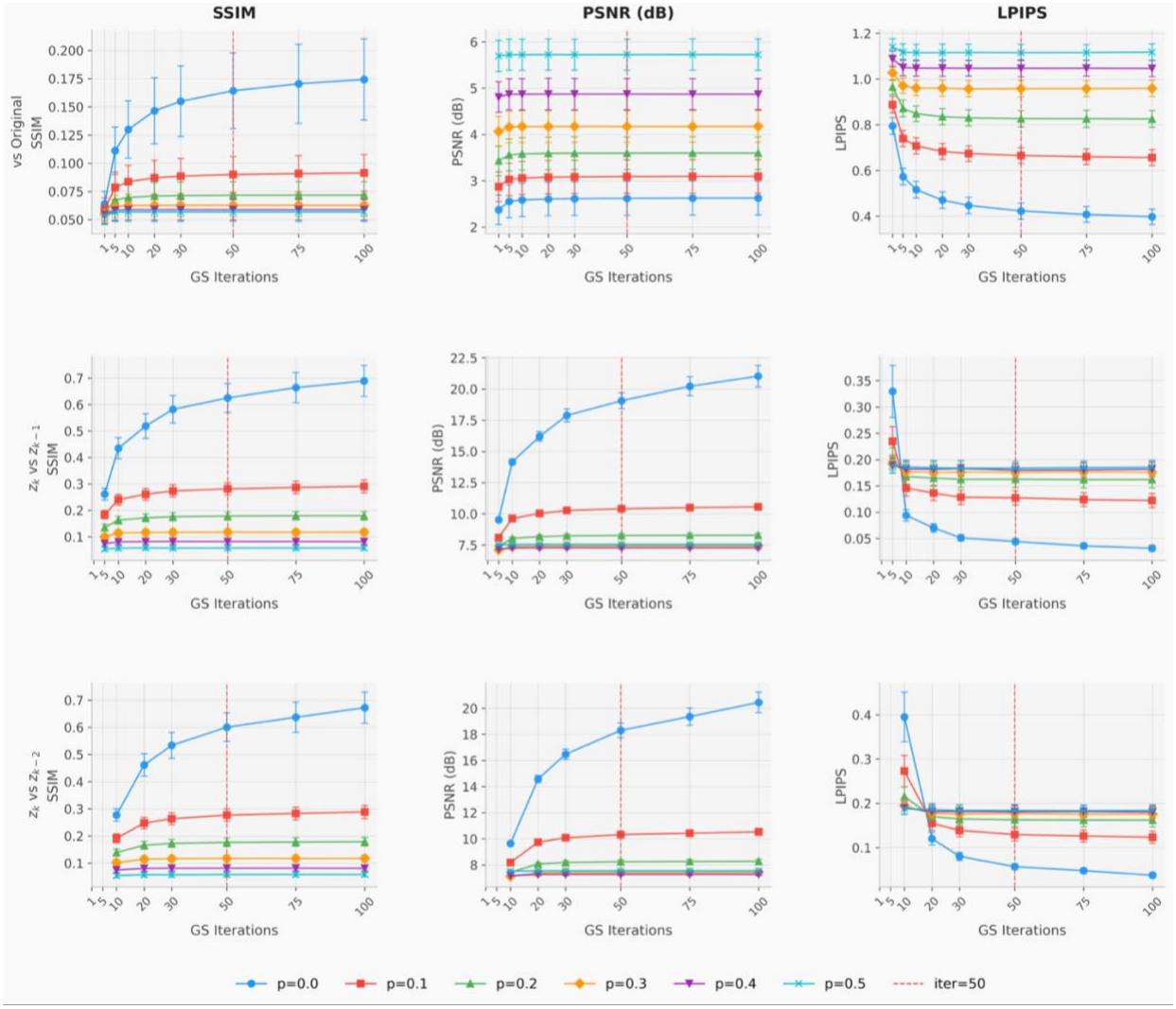
These analyses are not intended to establish formal privacy guarantees. Rather, they provide empirical justification for the transformation settings used throughout the main manuscript. In particular, they support the use of $I = 50$ iterations as a stable and computationally practical operating point and characterize how stochastic variability behaves when outputs are aggregated across images.

S2.2 GS Iteration Convergence and Optimization

The GS transformation used throughout this work is an iterative projection procedure whose behavior depends on both the number of iterations I and the masking probability p. In the main experiments, the number of iterations is fixed at $I = 50$. This choice reflects a practical trade-off between representational stability and computational cost rather than a theoretically optimal stopping criterion.

To empirically evaluate this choice, we analyzed GS convergence behavior across a fixed set of iteration counts, $I \in \{1,5,10,20,30,50,75,100\}$, and masking probabilities, $p \in \{0.0,0.1,0.2,0.3,0.4,0.5\}$. For each condition, GS-transformed outputs were generated and compared using three complementary views of convergence. First, we measured similarity between each GS-transformed output and the original image, which captures how the transformed representation changes relative to the raw anatomy as iteration count increases. Second, we measured similarity between outputs generated at consecutive evaluated iteration settings, denoted z_k and z_{k-1}. Here, k indexes the ordered iteration settings rather than individual algorithmic iterations; for example, z_k at $I = 50$ is compared with z_{k-1} at $I = 30$. Third, we measured similarity between outputs separated by two evaluated iteration settings, denoted z_k and z_{k-2}; for example, $I = 50$ is compared with $I = 20$. This provides a more conservative assessment of stability across broader iteration-count changes. For each comparison, SSIM, PSNR, and LPIPS were computed and aggregated across images.

GS Iteration Optimization (using OASIS Dataset)



Supplementary Figure S6. GS convergence was evaluated across iteration counts $I \in \{1,5,10,20,30,50,75,100\}$ and masking probabilities $p \in \{0.0,0.1,0.2,0.3,0.4,0.5\}$. Metrics are reported relative to the original image and between outputs generated at adjacent or near-adjacent evaluated iteration settings. Comparisons z_k versus z_{k-1} measure changes between consecutive evaluated iteration settings, while z_k versus z_{k-2} measures changes across a wider two-setting interval. The curves show rapid early change followed by stabilization around 30-50 iterations, supporting the use of $I = 50$ as a stable and computationally practical setting for the main experiments.

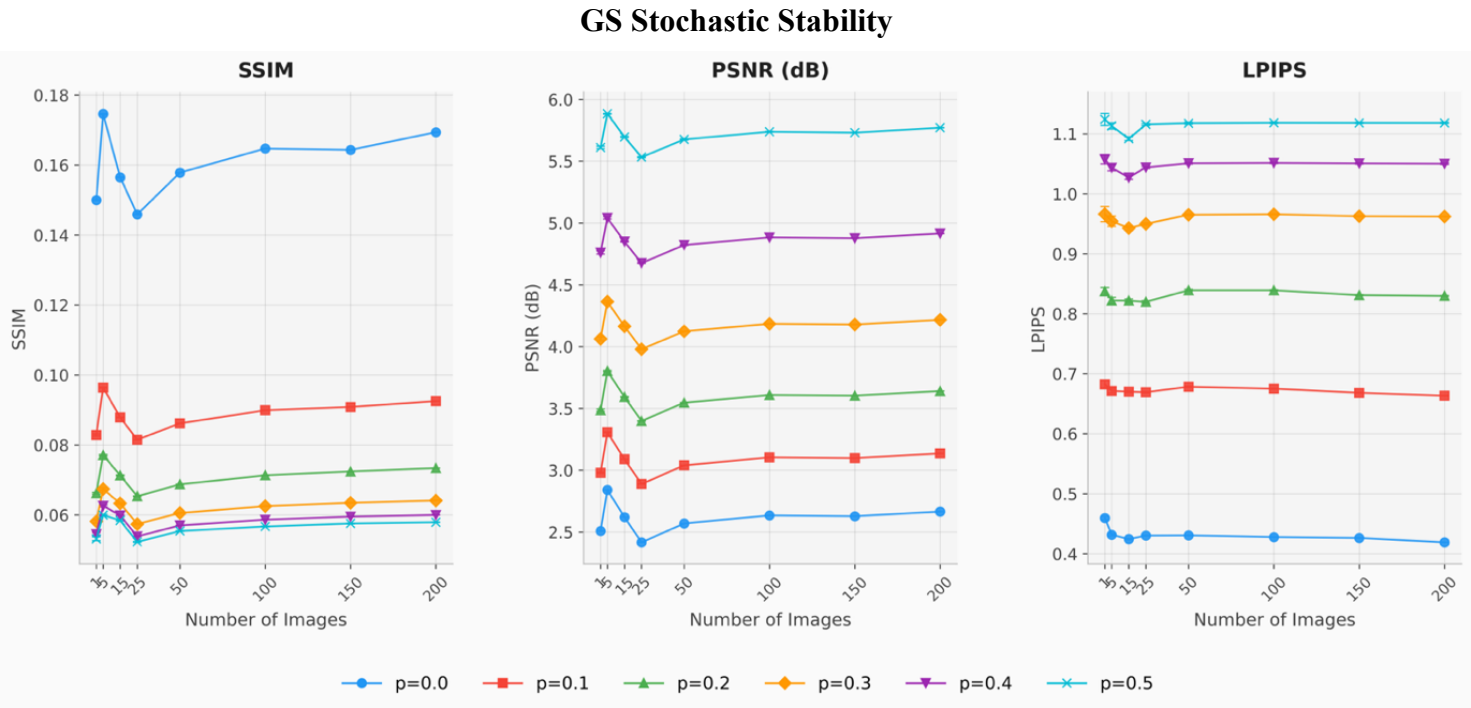
Supplementary Figure S6 summarizes the convergence analysis. Across masking strengths, metrics measured against the original image show rapid early change followed by diminishing differences as the iteration count increases. Inter-setting comparisons show the same general pattern: similarity between z_k and z_{k-1}, as well as between z_k and z_{k-2}, increases over early iteration settings and plateaus by approximately 30-50 iterations. Beyond this range, additional iterations produce only marginal changes in image-level convergence and inter-setting stability metrics.

These results support using $I = 50$ in the main experiments. This setting lies within the stabilized region of the iterative process across masking probabilities while avoiding unnecessary computational overhead from higher iteration counts. Thus, $I = 50$ should be interpreted as an empirically justified and computationally practical operating point, not as a theoretically optimal stopping rule.

S2.3 Stochastic Stability of GS Transformations

As described in Supplementary Material S1, the GS transformation incorporates stochastic phase initialization and per-iteration frequency masking. Therefore, repeated application of GS with the same input image and same algorithmic parameters can produce different pixel-level outputs unless the random initialization and masks are fixed. The purpose of the stochastic-stability analysis is to determine whether this image-level stochasticity meaningfully affects aggregate behavior at the dataset level.

For each masking level, GS transformations were computed using 10 independent random realizations at a fixed iteration count of $I = 50$. Reconstruction similarity between GS-transformed outputs and the original images was evaluated using SSIM, PSNR, and LPIPS. Metrics were aggregated over increasing sample sizes, $n \in \{1,5,15,25,50,100,150,200\}$, and variability across random realizations was quantified using the standard deviation.



Supplementary Figure S7. GS transformations were repeated using 10 independent random realizations at fixed $I = 50$ across masking probabilities $p \in \{0.0,0.1,0.2,0.3,0.4,0.5\}$. SSIM, PSNR, and LPIPS were computed between GS-transformed outputs and the corresponding original images, then aggregated across increasing sample sizes $n \in \{1,5,15,25,50,100,150,200\}$. Variability across random realizations is largest for very small sample sizes but decreases rapidly as more images are aggregated. The relative ordering of masking strengths remains stable across random realizations, indicating that GS stochasticity affects individual images but does not destabilize aggregate transformation behavior under the tested conditions.

Supplementary Figure S7 summarizes the stochastic-stability analysis. Variability across random realizations is most pronounced when metrics are computed over very small numbers of images. This is expected because individual GS outputs vary at the pixel level depending on the sampled phase field and masks. However, variability decreases rapidly as the number of aggregated images increases. By moderate sample sizes, approximately 25-50 images, mean metric values stabilize, and inter-realization variability becomes small across masking levels. Increasing the sample size beyond this range produces limited additional variance reduction.

The relative ordering of masking strengths is preserved across random realizations and sample sizes. Stronger masking consistently produces lower SSIM and PSNR and higher LPIPS relative to the original image. This indicates that stochastic sampling does not obscure the qualitative effect of masking strength

under dataset-level aggregation. Instead, image-level stochasticity averages out as more images are included.

These results support the interpretation that GS is stochastic at the individual-image level but statistically stable when evaluated over image sets. Within the tested conditions, the aggregate behavior of the transformation is therefore not driven by a single random initialization or mask realization, supporting the use of once-generated GS datasets in the main experiments.

S2.4 Computational Environment

All experiments were conducted in a cloud-based JupyterLab environment hosted on Google Cloud Vertex AI. The computational environment described here applies to the full experimental workflow, including GS transformation, model training, re-identification evaluation, reconstruction experiments, and runtime profiling. It is reported in this section because the GS runtime analysis depends directly on the same hardware and software configuration.

The compute instance was configured with 16 virtual CPUs, 60 GB RAM, and one NVIDIA Tesla P100 GPU. Experiments were run in a Vertex AI Deep Learning image using Ubuntu and JupyterLab. The primary software libraries included PyTorch, TensorFlow, NumPy, SciPy, scikit-learn, Matplotlib, and CuPy for GPU-accelerated GS transformation.

This environment was selected to provide a consistent execution platform across experiments while supporting GPU-accelerated spectral transformation and deep learning workflows. All runtime measurements reported in the following section were collected under this environment. Absolute runtimes may differ on other hardware, but the scaling trends with image size, batch size, and iteration count are expected to generalize.

Component	Configuration
Cloud platform	Google Cloud Vertex AI
Execution interface	JupyterLab
Operating system	Ubuntu-based Vertex AI Deep Learning image
CPU/RAM	16 vCPUs, 60 GB RAM
GPU	1 × NVIDIA Tesla P100
Region	us-east1-b
Primary deep learning libraries	PyTorch, TensorFlow
Primary numerical/scientific libraries	NumPy, SciPy, scikit-learn, Matplotlib
GPU array library for GS implementation	CuPy

Table S3. Computational Environment

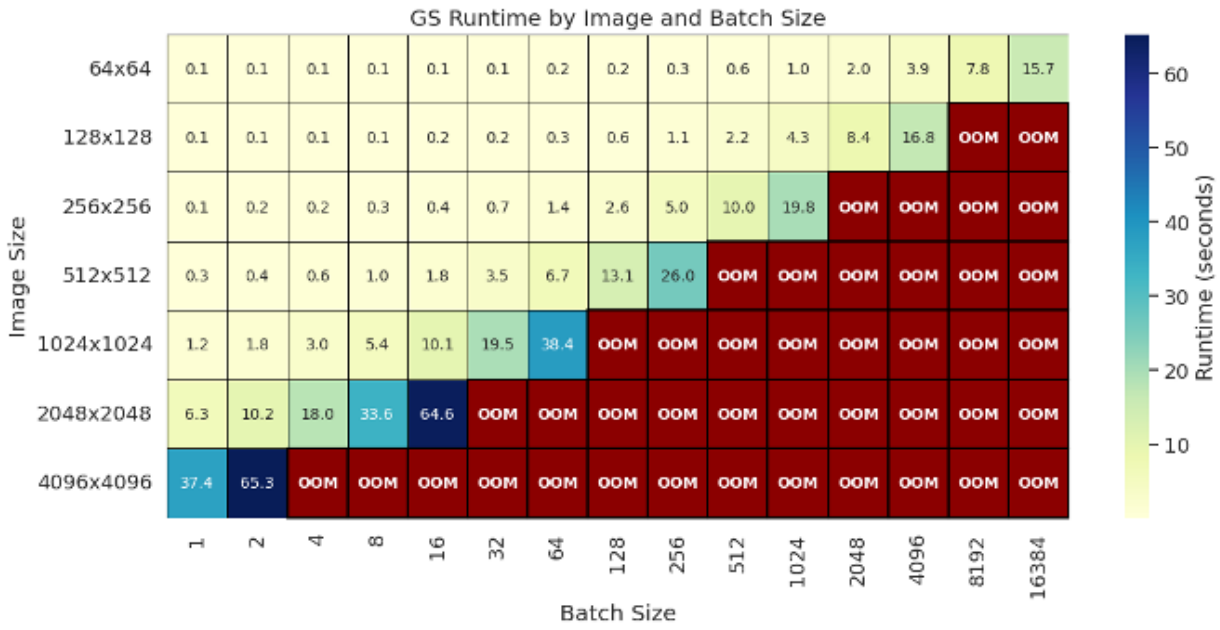
S2.5 Runtime Profiling of the GS Transformation

To characterize the computational cost of the GS transformation, runtime performance was benchmarked across input image resolutions and batch sizes. Image resolutions ranged from 64×64 to 4096×4096 pixels, and batch sizes ranged from 1 to 16,384 images per batch. For each configuration, the average wall-clock runtime per batch was measured using the fixed GS implementation and runtime environment described above. Configurations that exceeded available GPU memory were recorded as out-of-memory (OOM).

Supplementary Figure S8 summarizes the runtime analysis as a heatmap. Runtime increases with both image resolution and batch size, reflecting the cost of repeated Fourier-domain transformations, inverse transformations, complex-valued array operations, and GPU memory movement. High-resolution images impose the largest computational burden and lead to OOM failures at moderate or large batch sizes. In contrast, low- and medium-resolution images remain tractable across a wider range of batch sizes.

The masking probability p has minimal impact on runtime. Masking is implemented as element-wise multiplication in the frequency domain, which contributes negligible overhead compared with the dominant DFT, inverse DFT, and complex-valued matrix operations. Therefore, computational cost is governed

primarily by image resolution, batch size, and iteration count. These runtime results provide a practical reference for applying GS in resource-constrained settings. In the main experiments, using $I = 50$ balances convergence stability with computational feasibility. Higher iteration counts may be possible, but the convergence analysis in Supplementary Figure S6 indicates limited representational change beyond this range, while runtime cost increases with additional iterations.



Supplementary Figure S8. Runtime profiling was performed on the Google Cloud Vertex AI environment described in Table S3 using the GS implementation applied in this study. The heatmap reports average wall-clock runtime per batch across image resolutions from 64×64 to 4096×4096 pixels and batch sizes from 1 to 16,384 images. Configurations exceeding available GPU memory are marked as out-of-memory (OOM). Runtime increases with image resolution and batch size, while masking probability contributes negligible overhead because masking is implemented as element-wise multiplication relative to the dominant Fourier-domain operations.

Supplementary Material S3: Datasets, Cohort Construction, Preprocessing, and GS Dataset Generation

S3.1 Overview

This section documents the datasets, cohort construction, preprocessing, split design, and Gerchberg-Saxton (GS) dataset generation procedure used throughout the main experiments. The split designs differ across experiments because the underlying tasks differ. Diagnostic utility evaluation is open-set and uses subject-level disjoint splits. Supervised re-identification is closed-set and requires every subject to appear in train, validation, and test partitions, and is therefore split at the slice or image level. The corresponding cross-validation and bootstrap procedures are described per dataset in S3.2 and S3.3, with the cross-experiment summary in S3.5.

S3.2 OASIS-1

S3.2.1 Source and Label Scheme

Structural T1-weighted brain MRI scans were obtained from the public Kaggle mirror of OASIS-1 (`ninadaithal/imagesoasis`, Apache-2.0 license). The mirror provides axial slices at 224 × 224 resolution as preprocessed PNG images, organized by clinical-status label (Non-Demented, Very Mild Dementia, Mild Dementia, Moderate Dementia). The dataset comprises 347 subjects with axial brain MRI

slices in the range 100-160. Each subject contributes between one and six motion-corrected reconstruction (MPR-1 through MPR-6) versions per slice, with not all subjects having all six versions; the full file pool across all available MPR versions totals 86,437 slices (67,222 non-demented, 13,725 Very Mild, 5,002 Mild, 488 Moderate).

For binary dementia classification, the four clinical labels were collapsed into two: label 0 (Healthy) was assigned to non-demented subjects, and label 1 (Demented) was assigned to subjects in any of the three dementia categories.

S3.2.2 Cross-Experiment MPR Usage

The two OASIS-1 experimental tracks use different subsets of the available MPR versions:

- **Utility (S3.2.4):** All available MPR versions per subject are used. This maximizes per-subject sample count for the open-set classification task and reflects the typical usage pattern of the Kaggle release.
- **Re-identification (S3.2.5):** Only MPR-1 is used. MPR-1 is the version present for every subject across the eight slice positions used for re-identification, ensuring that all 347 subjects contribute the same eight slices under a closed-set protocol.
- **Reconstruction adversary fine-tuning (S3.2.6, applied only to DPS):** All available MPR versions of the slices in the re-identification train+val pool are used to support domain adaptation of the pretrained DPS model. Reconstruction *evaluation* uses MPR-1 only, matching the re-identification protocol.

This design ensures that re-identification and reconstruction-evaluation samples are consistent across all 347 subjects (MPR-1 only), while the utility task and the DPS fine-tuning step use the full available MPR pool to maximize training data.

S3.2.3 Slice-Band Selection (Utility Task)

For utility evaluation, a central slice band corresponding to slice indices 115-145 (inclusive) was retained. This band targets the dementia-relevant brain region while excluding peripheral slices dominated by non-discriminative content. After slice-band filtering across all MPR versions, 43,927 slices across 347 subjects (266 Healthy, 81 Demented) remained for the utility task.

S3.2.4 Utility Split (Subject-Level, 5-Fold Cross-Validation)

Subject-level partitioning was applied with no subject overlap across train, validation, and test. A held-out test set of 70 subjects (~20% of the cohort) was constructed first; the remaining 277 subjects form the 5-fold cross-validation pool, with 221 or 222 train subjects and 56 validation subjects per fold (the train count varies by one across folds because the CV pool size of 277 is not evenly divisible by five). The held-out test set is identical across all folds, so reported fold-level variability reflects training-data composition rather than test-set composition. Splits were stratified by binary label using random seed 1337, producing approximately balanced healthy-to-demented ratios per partition consistent with the cohort prevalence (266:81).

S3.2.5 Re-Identification Split (Slice-Level, Closed-Set)

The supervised re-identification task is closed-set: all 347 subjects appear in train, validation, and test partitions. Splitting is therefore performed at the slice level rather than the subject level. Each subject contributes 8 axial slices drawn from non-adjacent positions within the same MPR-1 acquisition: 100, 108, 115, 122, 130, 138, 145, and 153. These positions sit within the central slice band and are separated by 7 to 8 slice indices each, which is the maximum approximately uniform separation achievable across the 61-slice range when 8 positions must be drawn.

Test slice positions (122, 130, 138; 1,041 slices total) are held constant across all five cross-validation folds. The 5-fold rotation is applied to the train-validation boundary among the five remaining slice positions {100, 108, 115, 145, 153}, where one slice serves as validation in each fold and the remaining four serve

as training. This produces deterministic per-fold partition counts (1,388 train / 347 val / 1,041 test) that are constant across folds, with fold-level variance attributable to the rotating training-data composition.

Test slice positions are held fixed across folds for two methodological reasons. First, axial slice index corresponds to anatomical content; rotating test slices across folds would introduce variance from anatomical heterogeneity rather than from the GS transformation effect. Holding test slices constant ensures that fold-level differences in re-identification accuracy reflect classifier training variability under matched evaluation anatomy. Second, the closed-set protocol requires that all subjects appear in all three partitions, which is satisfied by this design at every fold.

S3.2.6 Reconstruction Inheritance and DPS Fine-Tuning

Reconstruction-based inversion experiments inherit the slice-level split from S3.2.5 directly. U-Net models are trained on paired raw-GS data using the train+val slices from MPR-1 only, and reconstructed outputs at the test slice positions (122, 130, 138) are evaluated with the re-identification classifiers from S3.2.5.

DPS *fine-tuning* additionally uses all available MPR versions of the train+val slices to support domain adaptation of the pretrained FFHQ-trained checkpoint, expanding the fine-tuning pool beyond the 1,735 MPR-1 train+val slices used by U-Net training. This expansion applies only to the fine-tuning data; evaluation slices remain MPR-1 at positions 122, 130, 138.

S3.3 NIH Chest X-ray

S3.3.1 Source and Class Filtering

Frontal chest radiographs were obtained from the public Kaggle release of the NIH ChestX-ray14 dataset (`nih-chest-xrays/data`). The full source contains 91,324 single-label images across 15 classes prior to filtering. Four classes were excluded due to insufficient per-subject representation, which prevents reliable subject-level splitting and balanced minimum-image filtering: Hernia (110 images), Pneumonia (322 images), Fibrosis (727 images), and Edema (628 images), and the remaining 11 are listed in Table S4.

For consistency across all NIH experiments, any subject whose label set included one or more of the four excluded classes was discarded from the cohort entirely, even when that subject also satisfied the minimum-image threshold for utility or re-identification eligibility. This ensures that the cohort used across utility and re-identification experiments is internally consistent, with no subject contributing data to one experiment while being excluded from another.

Index	Class	Source Count	Utility Train	Utility Val	Utility Test
0	No Finding	60,361	4,276	574	1,158
1	Infiltration	9,547	6,661	1,010	1,875
2	Atelectasis	4,215	2,955	446	813
3	Effusion	3,955	2,774	408	773
4	Nodule	2,705	1,916	298	491
5	Pneumothorax	2,194	1,481	253	459
6	Mass	2,139	1,507	231	401
7	Consolidation	1,310	944	146	220
8	Pleural Thickening	1,126	776	114	236
9	Cardiomegaly	1,093	739	131	223
10	Emphysema	892	625	114	153

Table S4. Eleven retained classes for the NIH Chest X-ray multi-label classification task with per-class counts in the source release and per-partition counts for the utility split (S3.3.2). Source counts reflect single-label images prior to subject-level filtering; partition counts reflect the filtered cohort.

S3.3.2 Utility Split (Subject-Level, Single Fixed Split)

For diagnostic utility evaluation, subjects with two or more frontal radiographs in the retained 11-class pool were selected, yielding 13,481 subjects and 35,181 images. Splits were constructed at the subject level with no patient overlap across partitions using random seed 1337, producing 9,436 train subjects (24,654 images), 1,346 validation subjects (3,725 images), and 2,699 test subjects (6,802 images). The split is fixed across experiments rather than rotated under cross-validation; confidence intervals are derived by subject-level bootstrap resampling (B=2000) over the test set, as described in S8. Per-class image counts are reported alongside the class list in Table S4.

S3.3.3 Re-Identification Split (Image-Level, Closed-Set, Single Fixed Split)

The supervised re-identification task is closed-set: all eligible patients appear in train, validation, and test partitions. Patient eligibility required at least five frontal radiographs in the filtered 11-class pool, yielding 1,739 patients and 10,038 total images. Splitting is performed at the image level: within each patient, three images are assigned to training, one image is assigned to validation, and any remaining images are assigned to testing. This rule produces 5,217 training images, 1,739 validation images, and 3,082 test images, with all 1,739 patients present in all three partitions. Test image counts vary across patients: 396 patients contribute exactly one test image (those with exactly five total images), and 1,343 patients contribute two or more test images. The split is fixed across experiments; confidence intervals are derived by subject-level bootstrap resampling (B=2000) over the test patients, as described in S8.

S3.3.4 Reconstruction Inheritance

The U-Net reconstruction adversary on chest X-ray inherits the re-identification split from S3.3.3 directly. Reconstruction models are trained on paired raw-GS data using the train+val images, and reconstructed outputs from test images are evaluated with the re-identification classifiers from S3.3.3. DPS reconstruction is not evaluated on chest X-ray (Section 3.3 of the main manuscript).

S3.4 Preprocessing

Both datasets enter the experimental pipeline as 224×224 single-channel grayscale images with pixel values normalized to $[0, 1]$. Preprocessing is minimal and applied prior to splitting and GS transformation.

For OASIS-1, the Kaggle mirror provides slices at 224×224 resolution. No additional preprocessing was applied beyond the slice-band filter for utility (S3.2.3), the MPR-1 selection for re-identification (S3.2.5), and verification that pixel values lie in $[0, 1]$. No skull stripping, bias-field correction, registration, or intensity windowing was applied.

For NIH Chest X-ray, images were resized to 224×224 with pixel values normalized to $[0, 1]$. No additional intensity normalization, clipping, or windowing was applied. Subject metadata (patient ID, study ID) was retained alongside the image arrays for split construction but was not used as model input.

After preprocessing, both datasets exist as float arrays of shape $(N, 224, 224)$ with values in $[0, 1]$. A single channel is added or removed at model-input time via `np.expand_dims` or `np.squeeze` according to the architecture's expected input shape; no other transformation is applied between preprocessing output and GS input.

S3.5 Cross-Experiment Split Summary

Table S5 consolidates the split designs across the four main experimental conditions. Embedding-based re-identification inherits the utility splits because gallery and probe sets are constructed from the held-out (or fixed test) subjects of the utility track. Reconstruction experiments inherit the supervised re-identification splits because reconstructed outputs are evaluated with the re-identification classifiers.

S3.6 GS Dataset Generation and Reuse Policy

GS-transformed datasets were generated once per (dataset, partition, masking level) combination prior to model training and cached for reuse across all downstream evaluations at the corresponding masking level. The generation procedure follows the order: (i) cohort filtering, (ii) preprocessing and intensity normalization to $[0, 1]$, (iii) construction of train/validation/test partitions, (iv) application of GS transformation to each partition independently. No post-GS preprocessing is applied; transformed images are passed to downstream models in the same numerical range and shape as their pre-transformation counterparts. A single channel dimension is inserted or removed at model-input time as required by the receiving architecture (S3.4).

Experiment	Dataset	Split Granularity	CV / Bootstrap	Subjects per partition (Train / Val / Test)	Images per partition (Train / Val / Test)
Diagnostic utility	OASIS-1	Subject	5-fold CV (held-out test, 70 subj.)	221-222 / 56 / 70	varies / varies / fixed
Diagnostic utility	NIH CXR	Subject	Single split, B=2000 bootstrap	9,436 / 1,346 / 2,699	24,654 / 3,725 / 6,802
Supervised re-ID	OASIS-1	Slice (MPR-1)	5-fold CV (fixed test slices)	347 / 347 / 347	1,388 / 347 / 1,041
Supervised re-ID	NIH CXR	Image	Single split, B=2000 bootstrap	1,739 / 1,739 / 1,739	5,217 / 1,739 / 3,082
Embedding-based re-ID	OASIS-1	Subject (inherits utility)	Inherits utility CV	inherited	inherited
Embedding-based re-ID	NIH CXR	Subject (inherits utility)	Inherits utility split + bootstrap	inherited	inherited
Reconstruction (U-Net)	OASIS-1	Slice (MPR-1, inherits re-ID)	Single fold (fold 0)	347 / 347 / 347	1,388 / 347 / 1,041
Reconstruction (U-Net)	NIH CXR	Image (inherits re-ID)	Single split	1,739 / 1,739 / 1,739	5,217 / 1,739 / 3,082
Reconstruction (DPS)	OASIS-1	Slice (MPR-1 eval; all-MPR fine-tune)	Single fold (fold 0)	347 / 347 / 347	1,388 / 347 / 1,041 (eval)

Table S5. Cross-experiment split summary. CV = cross-validation. Reconstruction models are trained on a single fold or split per dataset due to computational cost; reconstruction-output CIs follow the corresponding re-identification CI procedure (S8). DPS fine-tuning expands the OASIS-1 fine-tuning pool to all available MPR versions of the train+val slices while keeping evaluation slices at MPR-1.

Generation parameters follow Algorithm 1 (S1) and the parameter analysis in S2. The iteration count is fixed at $I = 50$ across all conditions, supported by the convergence and stability analyses in S2.2 and S2.3. Masking probability p is varied across the conditions reported in the main manuscript: utility, supervised re-identification, and embedding-based re-identification experiments use $p \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$; reconstruction experiments use $p \in \{0.0, 0.5\}$ as the weakest and strongest evaluated conditions.

GS transformations are applied independently per image, with no shared random state across images within a partition or across partitions. The random phase initialization (S1.5) and per-iteration binary mask draws (S1.7) are not seeded for individual GS generation runs, consistent with the stochastic-stability analysis in S2.3 which establishes that aggregate behavior is stable under independent random realizations once dataset-level aggregation is applied. Each (dataset, partition, p) combination is transformed exactly once and the result cached as a serialized array (pickle format) of shape matching the input partition; cached arrays preserve the original $(N, 224, 224)$ shape and float-valued pixel range.

Downstream experiments load the cached GS-transformed arrays directly. For each experimental condition reported in the main manuscript, the cached arrays for the corresponding (dataset, partition, p) combination are reused across cross-validation folds (where applicable), bootstrap resamples, and both architectures (ResNet18, DenseNet121). This ensures that observed differences across folds, bootstrap resamples, architectures, or evaluation conditions reflect the corresponding experimental factor rather than repeated stochastic GS sampling. The released code repository (Section 7 of the main manuscript) includes both the

cached-array workflow used for paper experiments and an on-the-fly fallback path for users who skip the cache-generation step; results presented in the main manuscript correspond to the cached-array workflow exclusively.

Supplementary Material S4: Diagnostic Utility Experiment Details

S4.1 Overview

This section documents the diagnostic utility experiment: the rationale for model configuration selection, architectural adaptations, training and evaluation procedures, and reproducibility settings. Hyperparameters and architectural specifics are consolidated in Table S6. The utility task is binary dementia classification on OASIS-1 and 11-class multi-label thoracic-pathology classification on NIH Chest X-ray. Cohort and split details are in S3; CI procedures are in S8.

S4.2 Model Configuration Selection (Pretrained vs. Scratch)

For each (dataset, architecture) pair, both ImageNet-pretrained and from-scratch training configurations were available. The main manuscript reports the configuration that achieved the highest raw-baseline performance per dataset, selected before evaluating GS transformations. This rule produced ImageNet-pretrained models with partial-layer fine-tuning for OASIS-1 and from-scratch models for NIH Chest X-ray.

This selection rule has two motivations. First, evaluating GS against the strongest available raw baseline provides a more conservative test of utility retention: any decline observed under GS is measured against the strongest non-GS baseline, rather than a weaker one that would inflate apparent retention. Second, the same utility models are reused for embedding-based re-identification (S6); higher-performing utility models are expected to produce more discriminative embeddings, providing a stronger upper-bound condition for embedding-based subject-identification risk. Reporting the higher-performing variant therefore strengthens both the utility-retention and the subject-identification-risk arms of the evaluation.

S4.3 Architectural and Training Configuration

Both ResNet18 and DenseNet121 were adapted for single-channel grayscale input by replacing the first convolutional layer (averaging the three pretrained input-channel weights for OASIS-1 pretrained models, or Kaiming-normal initialization for NIH scratch models) and replacing the classification head to match each task's output size, with task-specific loss, optimizer, and augmentation settings consolidated in Table S6.

OASIS-1 pretrained models used partial-layer fine-tuning: only the deeper feature blocks and the classification head were trainable, while earlier layers retained their ImageNet-pretrained weights. For ResNet18 this corresponds to the last two residual stages plus the head; for DenseNet121 this corresponds to the last two dense blocks plus the head. NIH Chest X-ray models trained all parameters from random initialization. Augmentation was applied only to NIH training because the OASIS-1 task operates on slices within a fixed central anatomical band where rotation, flipping, and intensity perturbation could disturb orientation cues that the classifier relies on.

S4.4 Evaluation Procedure

Diagnostic utility was evaluated using subject-level AUC under matched GS conditions (GS-to-GS training and evaluation). Per-image inference outputs were aggregated to subject-level scores by mean before AUC computation:

- **OASIS-1 (binary):** Per-image SoftMax P(y=1) was averaged across all evaluation slices of a subject to produce a single per-subject score. Subject-level AUC was computed across the test subjects against their binary labels.

- **NIH Chest X-ray (multi-label):** Per-image sigmoid outputs were averaged across each subject's images to produce per-subject per-class scores. Per-class AUC was computed across test subjects against their multi-hot labels, then macro-averaged across the 11 classes to produce the reported subject-level AUC.

The reported AUC values reflect a single training run per (dataset, architecture, GS level) combination. For OASIS-1, this single training run is repeated across the five cross-validation folds; the reported mean and 95% CI are computed across folds via t-interval (df=4). For NIH Chest X-ray, this single training run is performed on the fixed split (S3.3.2); the reported mean and 95% CI are computed via subject-level bootstrap resampling (B=2000) over the test subjects, applied to the single trained model's per-subject scores. Detailed CI formulations are in S8.

Component	OASIS-1 (pretrained)	NIH Chest X-ray (scratch)
Task	Binary classification (Healthy / Demented)	11-class multi-label classification
Input	224×224 single-channel, range [0, 1]	224×224 single-channel, range [0, 1]
Pretrained weights	ImageNet (ResNet18 / DenseNet121)	None
Trainable parameters (ResNet18)	Last two residual stages and head only	All parameters
Trainable parameters (DenseNet121)	Last two dense blocks and head only	All parameters
Classification head	2-output linear layer with dropout p=0.4	11-output linear layer with dropout p=0.5
Head initialization	Xavier-uniform weights, zero bias	Xavier-uniform weights, zero bias
Training loss	Class-weighted cross-entropy (balanced inverse-frequency) with label smoothing 0.05	Multi-label binary cross-entropy with per-class positive weighting (inverse frequency, clamped to [0.1, 20])
Validation loss	Cross-entropy, unweighted	Multi-label binary cross-entropy, unweighted
Inference activation	SoftMax (per-image class probabilities)	Per-class sigmoid (independent)
Optimizer	Adam	Adam
Learning rate	5×10^{-5}	8×10^{-5}
Weight decay	1×10^{-4}	5×10^{-4}
Batch size	128	128
Maximum epochs	50	100
Learning rate schedule	Reduce on plateau (×0.5 after 4 epochs without improvement, monitored on validation AUC)	Reduce on plateau (×0.5 after 4 epochs without improvement, monitored on validation AUC)
Early stopping	After 7 epochs without validation-AUC improvement; best-AUC checkpoint retained	After 10 epochs without validation-AUC improvement; best-AUC checkpoint retained
Gradient clipping	Maximum norm 1.0	Maximum norm 1.0
Mixed precision	Enabled (CUDA AMP)	Enabled (CUDA AMP)
Training augmentation	None	Horizontal flip (50%), rotation up to ±10°, brightness and contrast jitter ±20% (50%)
Validation/test augmentation	None	None

Table S6. Diagnostic utility training configuration for OASIS-1 and NIH Chest X-ray. Pretrained vs. scratch selection follows the highest-raw-baseline rule (S4.2). Pixel range entering the model is [0, 1] following the preprocessing pipeline in S3.4; no further input normalization is applied at the model-input stage.

S4.5 Reproducibility

All dataset splits used random seed 1337 (S3.2.4, S3.3.2). No further seed control (model initialization, `DataLoader` shuffling, augmentation, batch ordering) was applied unless explicitly specified above. Variability across the 5-fold CV (OASIS-1) and across bootstrap resamples (NIH Chest X-ray) therefore incorporates both data-partition or resample variability and run-level RNG variability inherent to unseeded model initialization and `DataLoader` shuffling. The released code repository (Section 7 of the main manuscript) specifies all hyperparameters as constants and reproduces the reported results when executed against the same dataset releases identified in S3.

Supplementary Material S5: Supervised Re-Identification Experiment Details

S5.1 Overview

This section documents the supervised re-identification experiments reported in Sections 3.1 and 3.2 of the main manuscript: cross-domain re-identification (classifier trained on GS-transformed images, evaluated on raw images) and in-domain re-identification (classifier trained and evaluated on GS-transformed images at the same masking level). Both conditions share the same architecture and training configuration; they differ only in the data domain at evaluation time. Cohort and split details are in S3; CI procedures are in S8.

S5.2 Experimental Conditions

The cross-domain and in-domain conditions probe distinct subject-identification risks under GS transformation.

Cross-domain re-identification. A closed-set classifier is trained on GS-transformed slices (or images, for NIH) at masking level p and evaluated on the corresponding raw test partition without any GS transformation applied at evaluation. This condition measures whether identity-related features learned in the GS-transformed domain transfer back to raw images.

In-domain re-identification. A closed-set classifier is trained and evaluated entirely within the same GS-transformed domain at masking level p. This condition establishes an upper bound on how much identity-bearing structure remains learnable within the transformed representation itself, independent of any domain boundary crossing.

Component	OASIS-1	NIH Chest X-ray
Task	Closed-set subject classification	Closed-set subject classification
Output classes	347 subjects	1,739 patients
Input	224×224 single-channel, range [0, 1]	224×224 single-channel, range [0, 1]
Pretrained weights	ImageNet (ResNet18 / DenseNet121)	ImageNet (ResNet18 / DenseNet121)
Trainable parameters	All parameters	All parameters
First-layer initialization (pretrained)	Luminosity-weighted RGB collapse (0.299, 0.587, 0.114)	Luminosity-weighted RGB collapse (0.299, 0.587, 0.114)
Classification head	Linear layer to 347 classes, no dropout	Linear layer to 1,739 classes, no dropout
Training loss	Cross-entropy with label smoothing 0.1	Cross-entropy with label smoothing 0.1
Class weighting	None	None
Optimizer	AdamW	AdamW
Learning rate (peak)	5×10^{-4}	5×10^{-4}
Weight decay	1×10^{-4}	1×10^{-4}
Batch size	16	64
Maximum epochs	30	30
Learning rate schedule	Cosine decay with linear warmup over the first 10% of total training steps; per-batch update	Cosine decay with linear warmup over the first 10% of total training steps; per-batch update
Early stopping	After 8 epochs without validation Top-1 improvement; best-Top-1 checkpoint retained	After 8 epochs without validation Top-1 improvement; best-Top-1 checkpoint retained
Gradient clipping	Maximum norm 1.0	Maximum norm 1.0
Mixed precision	Enabled (CUDA AMP)	Enabled (CUDA AMP)
Per-image input normalization	Per-split mean and standard deviation, computed from training images of that GS condition	Per-condition mean and standard deviation, computed from pooled train + val images of that GS condition
Training augmentation	Random rotation up to ±5° (50%), horizontal flip (50%), brightness scaling within ±20% (50%)	Same as OASIS-1
Validation/test augmentation	None	None

Table S7. Supervised re-identification training configuration. Both datasets use ImageNet-pretrained models with full fine-tuning and a closed-set classification head sized to the subject count. Differences in batch size reflect available memory under each input resolution and per-batch class count; other configuration values are shared across datasets to keep cross-dataset comparisons under the same training protocol.

A separate classifier is trained for each (dataset, architecture, p, condition) combination. The raw-baseline row in Tables 1 and 2 of the main manuscript corresponds to a classifier trained and evaluated on raw images and is identical between cross-domain and in-domain conditions.

S5.3 Architectural and Training Configuration

Both ResNet18 and DenseNet121 are adapted for single-channel grayscale input as in the utility experiments (S4.3), with a classification head sized to the closed-set subject count (347 for OASIS-1, 1,739 for NIH Chest X-ray). The complete configuration is given in Table S7.

The supervised re-identification configuration differs from the utility configuration in several methodologically motivated ways. Re-identification uses a much larger output space (347 or 1,739 classes vs. 2 or 11), no class weighting (all subject classes are intentionally balanced under the closed-set protocol), and full fine-tuning of all parameters rather than partial-layer fine-tuning (the task requires learning subject-discriminative features that are not well-aligned with the partial ImageNet features retained by partial fine-tuning). The configuration follows a single-rule selection: ImageNet-pretrained models with full fine-tuning, selected for the strongest baseline re-identification performance under the rule described in S4.2, for re-identification and reconstruction-related models, the strongest-baseline rule is applied to identify the configuration most likely to expose residual subject-identifying signal under GS, providing a conservative upper bound on subject-identification risk.

S5.4 Cross-Domain vs. In-Domain Training and Evaluation

The architecture and hyperparameters in Table S7 are identical for both conditions. The only difference is the data presented at evaluation time:

Cross-domain protocol. Training set: GS-transformed images at masking level p. Validation set: GS-transformed images at the same masking level. Test set: raw (untransformed) images. Per-image normalization at test time uses the mean and standard deviation computed from the GS-transformed training data, not from the raw test data, so that the raw test inputs are normalized using statistics consistent with the training distribution.

In-domain protocol. Training, validation, and test sets are all GS-transformed at the same masking level p. Per-image normalization uses the mean and standard deviation computed from the GS-transformed training data of that masking level.

This design isolates the domain shift to a single factor: evaluation-time data domain. Architecture, training procedure, training data, and normalization policy are held constant across the two conditions.

S5.5 Evaluation Procedure

Re-identification performance is reported as Top-1 accuracy: the fraction of test images for which the classifier's highest-probability output corresponds to the correct subject identity. Top-1 accuracy is computed at the image level (no per-subject aggregation) because the closed-set protocol assigns each test image a single ground-truth subject label, and re-identification is the per-image classification task itself.

Chance-level Top-1 accuracy under uniform random guessing is 1/347 = 0.288% for OASIS-1 and 1/1,739 = 0.058% for NIH Chest X-ray.

For OASIS-1, mean and 95% CI are computed across the 5-fold cross-validation results via t-interval (df = 4). For NIH Chest X-ray, mean and 95% CI are derived by subject-level bootstrap resampling (B = 2000) over the test patients of the single fixed split: each bootstrap resample draws 1,739 patients with replacement, recomputes per-patient accuracy from the trained classifier's predictions on the resampled patients' test images, and aggregates into a Top-1 accuracy estimate. The 2.5th and 97.5th percentiles of the resulting distribution define the 95% CI. Detailed CI formulations are in S8.

S5.6 Reproducibility

All dataset splits use random seed 1337 (S3.2.5, S3.3.3). The OASIS-1 5-fold CV uses deterministic slice-position rotation (S3.2.5), fold index k assigns slice index from the train+val pool {100, 108, 115, 145, 153} at position k as validation and the remaining four as training; no additional seed is required for fold construction. The NIH Chest X-ray re-identification split is fixed and seeded once at construction.

Per-image normalization statistics (mean, standard deviation) are recomputed from each (dataset, GS level, partition) combination as a function of the data, with no separate seed control. Beyond the split seed, no further seed control is applied to model initialization, DataLoader shuffling, or augmentation, consistent with the policy in S4.5; variability across the 5-fold CV (OASIS-1) and across bootstrap resamples (NIH Chest X-ray) therefore incorporates both data-level and run-level RNG variability.

The released code repository (Section 7 of the main manuscript) specifies all hyperparameters as constants and reproduces the reported results when executed against the same dataset releases identified in S3.

Supplementary Material S6: Embedding-Based Re-Identification Details

S6.1 Overview

This section documents the embedding-based re-identification experiment reported in Section 3.2 of the main manuscript. Embedding-based re-identification reuses the utility-trained classifiers from S4 as feature extractors; penultimate-layer activations are used as identity embeddings, and re-identification is performed by cosine-similarity matching between probe and gallery embeddings under matched GS conditions. This experiment is open-set in subject membership (gallery and probe sets are constructed from the held-out test subjects of the utility split, S3.2.4 and S3.3.2) and matched-domain in GS condition (gallery and probe are transformed at the same masking level). Cohort and split details are in S3; CI procedures are in S8.

S6.2 Embedding Extraction

Embeddings are extracted from the penultimate layer of the utility-trained model, the spatial-pooling output that immediately precedes the classification head. For ResNet18 the embedding is the flattened output of avgpool (512 dimensions); for DenseNet121 it is the flattened output of the global average pool over features → ReLU (1024 dimensions). The dropout layer and final linear classifier of the utility head are bypassed. The model is set to evaluation mode before extraction.

Each extracted embedding is L2-normalized per vector immediately after extraction. Extraction is performed in batches of 128 images.

S6.3 Subject-Level Aggregation

For each subject, slice-level (OASIS-1) or image-level (NIH Chest X-ray) embeddings are aggregated into a single subject-level embedding by computing the mean over the subject's images, then re-normalizing the resulting mean vector to unit L2 norm. This aggregation produces one identity embedding per subject for both gallery and probe sets, consistent with the per-subject reporting used elsewhere in the manuscript.

S6.4 Gallery and Probe Construction

Gallery and probe sets are constructed from the held-out test subjects of the utility split (S3.2.4 for OASIS-1, S3.3.2 for NIH Chest X-ray). For each subject in the test set, available images are partitioned deterministically into gallery and probe subsets using a per-subject seeded shuffle derived from the global seed 1337.

The default allocation rule assigns half of each subject's images to the gallery (rounded down, with a minimum of one image per side) and the remainder to the probe set. After this initial split, gallery size is capped at the configured value k. The main manuscript reports k = 1 and k = 5; the probe size cap is fixed at 1 image per subject.

For NIH Chest X-ray, only patients with at least 2 images in the held-out test pool are retained, so that each retained patient contributes both gallery and probe entries. For OASIS-1, all test subjects satisfy this requirement by construction (multiple slices per subject). No subject overlap exists between gallery and probe; the same subject set appears on both sides under the matched-domain protocol.

S6.5 Similarity Matching and Retrieval

Identity matching is performed by cosine similarity between probe and gallery embeddings, computed as the matrix product of L2-normalized embedding matrices. For each probe, the top-ranked gallery subject (highest cosine similarity) is taken as the predicted identity. Rank-1 accuracy is the fraction of probes for which the top-ranked gallery subject is the correct subject. Rank-5 accuracy is the fraction for which the correct subject appears among the five highest-ranked gallery entries.

The main manuscript reports Rank-1 accuracy at $k = 1$ and $k = 5$. Rank-5 accuracy and verification metrics (AUC, TAR at FAR=1%, mean positive and negative similarities) are computed and logged in the released code but are not reported in the main manuscript.

S6.6 Inheritance from Utility Models

The utility model used for embedding extraction is the same checkpoint reported as the utility classifier in Section 3.4 of the main manuscript (S4): ImageNet-pretrained ResNet18 or DenseNet121 with partial-layer fine-tuning for OASIS-1, and from-scratch ResNet18 or DenseNet121 for NIH Chest X-ray. For OASIS-1, the embedding evaluation inherits the 5-fold cross-validation structure of the utility experiment: a separate utility model is loaded per fold, embeddings are computed from each fold's model on its corresponding held-out test subjects, and Rank-1 accuracy is computed per fold. For NIH Chest X-ray, a single utility model trained on the fixed split is used for embedding extraction over the fixed held-out test subjects.

S6.7 Evaluation Procedure and Confidence Intervals

For OASIS-1, the reported mean and 95% CI are computed across the five folds via t-interval (df=4), with each fold contributing one Rank-1 accuracy value at each (architecture, GS level, k) combination. For NIH Chest X-ray, the reported mean and 95% CI are computed by subject-level bootstrap resampling (B=2000) over the test patients of the fixed split. Each bootstrap resample draws 1,739 test patients with replacement, recomputes Rank-1 accuracy from the trained model's embedding outputs on the resampled patients, and aggregates into a Rank-1 accuracy estimate. The 2.5th and 97.5th percentiles of the resulting distribution define the 95% CI. Detailed CI formulations are in S8.

S6.8 Reproducibility

Per-subject gallery/probe shuffling uses a deterministic per-subject random number generator seeded from the global seed 1337 combined with the subject identifier hash, ensuring that gallery/probe assignments are reproducible across runs. No further seed control is applied beyond this and the split seed (S3); model initialization and DataLoader-related randomness during utility training carries through to the embedding evaluation as documented in S4.5. The released code repository (Section 7 of the main manuscript) reproduces gallery/probe assignments, embeddings, and Rank-1 accuracies under the documented seed.

Supplementary Material S7: Reconstruction-Based Inversion Details

S7.1 Overview

This section documents the reconstruction-based inversion experiment reported in Section 3.3 of the main manuscript. Two adversary classes are evaluated: a U-Net trained from scratch to invert paired raw-GS data (S7.2) and a Diffusion Posterior Sampling (DPS) adversary using a pretrained diffusion model with optional domain-specific fine-tuning (S7.3). Both reconstruction adversaries operate at the input level, producing reconstructed images that are subsequently evaluated by the supervised re-identification classifiers from S5. The reconstruction adversary is trained or fine-tuned with direct exposure to the raw anatomical appearance of all evaluation subjects via paired raw-GS data, with only specific test slices or images

withheld from training, providing a strong upper-bound condition for reconstruction-based identity recovery.

S7.2 U-Net Reconstruction Adversary

S7.2.1 Purpose and Training Data

The U-Net is trained as a supervised image-to-image inversion network: input is a GS-transformed image, and target is the corresponding raw image. Separate models are trained at GS-0% and GS-50%, the weakest and strongest evaluated masking levels in the main manuscript.

For OASIS-1, training data is sourced from the supervised re-identification slice-level split (S3.2.5). The full image pool of 2,776 slices (1,388 train, 347 val, 1,041 test from the re-identification split) is re-partitioned at the subject level into a 70/10/20 train/val/test split with seed 1337 for U-Net training, with the additional constraint that the three reserved evaluation slice positions (122, 130, 138) are excluded from U-Net training and validation and are used only at U-Net test time. This reserved-slice exclusion prevents identity leakage from the U-Net training set into the slices used for downstream re-identification evaluation.

For NIH Chest X-ray, training data is sourced from the supervised re-identification image-level split (S3.3.3) directly: U-Net training uses the train and validation partitions, and the test partition is reserved for evaluation. No additional re-partitioning is applied because the NIH re-identification split is itself image-level and does not require reserved-slice exclusion.

S7.2.2 U-Net Architecture

The architecture is a standard 2D U-Net with skip connections, identical for both datasets. The encoder consists of three convolutional blocks at successively reduced spatial resolutions (32, 64, and 128 channels) followed by a 256-channel bottleneck, with 2 times max-pooling between levels. The decoder mirrors the encoder via three transposed-convolution up sampling stages, each followed by a convolutional block; skip connections concatenate encoder features with corresponding decoder features along the channel dimension. The output is produced by a 1×1 convolution mapping to a single channel followed by a sigmoid activation. Each convolutional block consists of two 3×3 convolutions with padding 1, each followed by an 8-group GroupNorm and SiLU activation. Input and output are single-channel 224×224 images with pixel values in [0, 1].

S7.2.3 U-Net Training Configuration

Training settings are largely shared between the two datasets, with adjustments for dataset size. Both use AdamW optimization with learning rate 2×10^{-4} and weight decay 1×10^{-4}, batch size 16, gradient clipping at maximum norm 1.0, and CUDA mixed precision. The combined training loss is L_1 reconstruction loss with weight 1.0 plus a $1 - \text{SSIM}$ term with weight 0.1, where SSIM is computed using a Gaussian window of size 11 and $\sigma = 1.5$. PSNR is logged as a tracking metric but does not enter the loss.

OASIS-1 U-Net training uses a maximum of 10 epochs with early stopping after 3 epochs without validation-loss improvement, reflecting the smaller training pool size (under 2,000 slices per partition) and the architecture's tendency to converge quickly on this scale of data. NIH Chest X-ray U-Net training uses a maximum of 50 epochs with early stopping after 5 epochs without validation-loss improvement, reflecting the larger training pool. Both datasets use the same improvement threshold of 1×10^{-5} decrease in validation loss and retain the best-validation-loss checkpoint. Determinism settings (`cudnn.deterministic=True, cudnn.benchmark=False`) are enabled in both cases. Input pixel values are passed directly in the [0, 1] range with no further per-image normalization beyond what is applied in the GS preprocessing pipeline (S3.4).

S7.2.4 Reconstruction Evaluation

At test time, the trained U-Net produces reconstructed test images from the GS-transformed test inputs. Reconstruction quality is reported via SSIM and PSNR computed against the corresponding raw test

images, with image-level bootstrap resampling (B=2000) over the test set providing 95% CIs. Reconstructed images are then passed to the supervised re-identification classifier from S5 (matched dataset and architecture), and per-image Top-1 accuracy is computed and reported. Re-identification CIs follow the dataset-specific procedures: 5-fold-CV t-interval for OASIS-1 (with reconstruction trained on fold 0 only and evaluated against all 5-fold-trained classifiers, as described in Section 2.5 of the main manuscript), and subject-level bootstrap (B=2000) for NIH Chest X-ray.

S7.3 Diffusion Posterior Sampling (DPS) Reconstruction Adversary

S7.3.1 Pretrained Model and Inference Framework

The DPS adversary uses the pretrained diffusion model and inference framework released by Chung et al. [18] from the public repository at https://github.com/DPS2022/diffusion-posterior-sampling. The pretrained checkpoint ffhq_10m.pt (FFHQ face-image diffusion model) is used as the base model. The base architecture is a 256×256-resolution score-based diffusion model with 128 base channels, residual blocks at each spatial level, attention at the 16×16 resolution, and 1,000 diffusion steps under a linear noise schedule. The released model has approximately 93.6 million parameters, of which the output blocks (output_blocks.6 through output_blocks.11) and the final output projection are designated trainable for fine-tuning (≈8.5 million parameters, 9.1% of total); all other layers are frozen.

S7.3.2 OASIS-Only Evaluation Rationale

DPS is evaluated on OASIS-1 only. OASIS-1 retains the higher residual identity signal under direct re-identification across both supervised and embedding-based conditions reported in the main manuscript (Sections 3.1 and 3.2), making it the more conservative test setting for evaluating whether a stronger reconstruction adversary can recover identity-bearing structure under GS transformation. Establishing that even a strong domain-adapted generative prior fails to recover subject-specific structure on the more resistant modality provides the upper-bound condition for reconstruction-based identity recovery, with computational cost for fine-tuning per modality as a supporting consideration. This rationale is summarized in Section 3.3 of the main manuscript.

S7.3.3 Domain-Specific Fine-Tuning

The pretrained FFHQ checkpoint is fine-tuned on OASIS-1 brain MRI to support domain adaptation. Fine-tuning data consists of 82,186 OASIS-1 axial slices drawn from all available MPR versions of the train+validation slices in the re-identification pool (S3.2.6), expanding the fine-tuning data beyond the MPR-1-only pool used for re-identification training and evaluation. Slices are resized from their original resolution to 256×256 via bilinear interpolation (align_corners=False), grayscale slices are repeated across three channels to match the input shape expected by the FFHQ-trained model, and pixel values are rescaled from [0, 1] to [−1, 1] for diffusion-model training.

Fine-tuning uses AdamW with learning rate 1×10^{-5} and weight decay 0, gradient clipping at maximum norm 1.0 over the trainable parameters, batch size 2, and a total of 2,000 training steps. The training objective is the standard simple noise-prediction loss used in denoising diffusion models, with uniform timestep sampling from 1,000 diffusion steps. The fine-tuned checkpoint is saved at the step of best training loss; the best loss obtained was 0.0032 at step 1,950.

S7.3.4 DPS Inference

At inference, posterior sampling is performed per image, with the initial sample drawn as standard Gaussian noise of the same shape as the reference image. Reverse diffusion sampling follows the framework's default configuration (provided via the repository's task and diffusion configuration YAML files, which are reproduced in the released code). Output images are de-rescaled from [−1, 1] back to [0, 1] before evaluation, with a clamp applied to keep values in valid range.

S7.3.5 DPS Reconstruction Evaluation

Reconstructed test images are evaluated using the same SSIM and PSNR procedure as the U-Net adversary (S7.2.4) and the same supervised re-identification classifier from S5. Both pretrained-only and fine-tuned DPS configurations are evaluated and reported in Table 3 of the main manuscript. The fine-tuned configuration produces anatomically coherent brain reconstructions reflecting successful domain adaptation, while the pretrained-only configuration produces outputs reflecting the FFHQ pretraining prior.

S7.4 Single-Run Reconstruction Training

Reconstruction models (both U-Net and DPS fine-tuning) are trained on a single configuration per dataset due to the substantial computational cost of training inversion networks at scale across multiple folds or seeds. For OASIS-1, U-Net is trained on the fold-0 partition; for NIH Chest X-ray, U-Net is trained on the single fixed re-identification split. DPS fine-tuning is performed once on OASIS-1. The resulting reconstructed test images are evaluated against all corresponding fold-trained or single-run re-identification classifiers from S5, with re-identification CIs computed under the dataset-specific procedure (OASIS: t-interval across the 5-fold-trained classifiers' evaluations of the fold-0 reconstructions; NIH: subject-level bootstrap over the test patients of the single split). This procedure provides classifier-side variability while reconstruction-model variability remains unestimated; this is acknowledged as a limitation in the main manuscript.

S7.5 Reproducibility

All splits and partitions inherited from S3 use the documented seed 1337. U-Net training uses seed 1337 with `cudnn.deterministic=True` and `cudnn.benchmark=False`; DPS fine-tuning operates on the documented data and steps without explicit seed control beyond the global seed inherited at process start. The released code repository (Section 7 of the main manuscript) reproduces both U-Net and DPS reconstruction outputs and downstream re-identification evaluations under the documented configurations.

Supplementary Material S8: Statistical Analysis and Confidence Intervals

S8.1 Overview

This section documents the confidence interval procedures used throughout the main manuscript. Two procedures are used, applied per dataset rather than per experiment: 5-fold cross-validation t-interval for OASIS-1, and subject-level bootstrap resampling for NIH Chest X-ray. Reconstruction-output SSIM and PSNR confidence intervals use an additional image-level bootstrap procedure documented in S8.5. Sections 2.5 and the table and figure captions of the main manuscript reference this section directly.

S8.2 5-Fold Cross-Validation t-Interval (OASIS-1)

For each (architecture, GS level, condition) combination, a metric value is computed per fold, yielding five values m_1, m_2, ..., m_5. The reported mean and 95% confidence interval are computed via Student's t-interval with four degrees of freedom:

$$\bar{m} = \frac{1}{5}\sum_{k=1}^{5} m_k, s = \sqrt{\frac{1}{4}\sum_{k=1}^{5}(m_k - \bar{m})^2}$$

$$95\%\ \text{CI} = [\bar{m} - t_{0.975,4} \cdot \frac{s}{\sqrt{5}}, \bar{m} + t_{0.975,4} \cdot \frac{s}{\sqrt{5}}]$$

where $t_{0.975,4} \approx 2.776$ is the 97.5th percentile of Student's t-distribution with four degrees of freedom. This procedure applies to all OASIS-1 utility AUC, supervised re-identification Top-1 accuracy, and embedding-based re-identification Rank-1 accuracy values reported in Tables 1, 2, and 4 of the main manuscript.

S8.3 Subject-Level Bootstrap (NIH Chest X-ray, Utility AUC)

For NIH Chest X-ray utility, a single trained model is evaluated on the 2,699 test patients. Confidence intervals are computed by subject-level bootstrap resampling with B = 2000 resamples:

1. For each bootstrap resample b ∈ {1, ..., 2000}, draw 2,699 patient indices uniformly at random with replacement from the test patient set.
2. From the trained model's per-image sigmoid outputs, compute per-patient per-class scores by averaging within each resampled patient (S4.4); duplicated patients contribute their per-patient scores once per occurrence in the resample.
3. Compute per-class AUC across the resampled patient set against multi-hot labels, then macro-average across the 11 classes to produce the resample's subject-level AUC value.
4. After all B resamples, the reported mean is the average of bootstrap AUC values, and the 95% CI is given by the 2.5th and 97.5th percentiles of the bootstrap distribution.

S8.4 Subject-Level Bootstrap (NIH Chest X-ray, Re-Identification)

For NIH Chest X-ray supervised and embedding-based re-identification, the same subject-level bootstrap procedure is applied with B = 2000 resamples over the 1,739 test patients of the closed-set re-identification split:

1. For each bootstrap resample, draw 1,739 patient indices with replacement.
2. From the trained classifier's predictions on the test images of each resampled patient, compute the patient's per-image Top-1 accuracy or Rank-1 accuracy, then average across the resampled patients (with duplicates counted per occurrence) to produce the resample's accuracy estimate.
3. The reported mean and 95% CI are derived from the 2,000 resample accuracies as in S8.3.

For embedding-based re-identification, gallery and probe sets are constructed once on the original test patient set (S6.4) and held fixed across bootstrap resamples; resampling reweights the contribution of each patient's probe-side accuracy without re-shuffling gallery/probe assignments, preserving the matched-domain protocol.

S8.5 Image-Level Bootstrap (Reconstruction SSIM and PSNR)

For reconstruction quality evaluation (Table 3), SSIM and PSNR confidence intervals are computed by image-level bootstrap resampling over the test images of the relevant split:

1. For each bootstrap resample b ∈ {1, ..., 2000}, draw N test-image indices with replacement, where N is the test-set size for the dataset (1,041 for OASIS-1; 3,082 for NIH Chest X-ray).
2. Compute the mean SSIM and mean PSNR across the resampled images.
3. The reported mean and 95% CI are derived from the 2,000 bootstrap means as in S8.3.

The image-level resampling unit is appropriate for SSIM and PSNR because these are per-image quality metrics, and within-subject correlation has limited bearing on perceptual reconstruction fidelity at the test-set distribution level. Re-identification accuracies on reconstructed images, by contrast, follow the dataset-specific procedures in S8.2 and S8.4.

S8.6 Reconstruction Re-Identification CI for OASIS-1

For OASIS-1 reconstruction-based re-identification (Table 3, Re-ID columns), reconstruction is performed once on fold-0 reconstructed test images, and the reconstructed images are evaluated against each of the 5-fold-trained re-identification classifiers from S5. The five resulting Top-1 accuracy values are aggregated via the t-interval (df=4) procedure of S8.2. This CI captures classifier-training variability across folds while reconstruction-model variability is not separately estimated, as discussed in S7.4.

S8.7 Hypothesis Testing

No formal pairwise hypothesis tests across GS masking levels are reported in the main manuscript. Confidence intervals overlap and absolute differences between GS levels are reported and discussed; the consistency of trends across architectures, datasets, and adversary classes is the primary basis for the main empirical claims rather than individual pairwise significance.

S8.8 Chance-Level Computation

For closed-set supervised re-identification, the chance-level Top-1 accuracy under uniform random guessing is 1/N, where N is the number of subjects in the closed-set classifier head: 1/347 = 0.288% for OASIS-1 and 1/1,739 = 0.0575% (rounded to 0.06% in the main manuscript) for NIH Chest X-ray. For diagnostic utility AUC, the chance level under random predictions is 0.5; reported values substantially above this baseline confirm that the trained classifiers have learned task-relevant signal.

Supplementary Material S9: Reproducibility Checklist and Output Mapping

S9.1 Software Environment

All experiments were executed in the cloud-based JupyterLab environment described in S2.4, with the libraries listed in Table S3 (NumPy, SciPy, scikit-learn, PyTorch, TensorFlow, CuPy, Matplotlib). Specific library versions used are pinned in the released code repository (Section 7 of the main manuscript). The runtime environment, hardware (NVIDIA Tesla P100, 16 vCPUs, 60 GB RAM), and operating system match the description in S2.4 across all experiments.

S9.2 Random Seeds

All dataset splits across OASIS-1 and NIH Chest X-ray use random seed 1337 for reproducible partitioning (S3.2.4, S3.2.5, S3.3.2, S3.3.3). Per-subject gallery/probe shuffling for embedding-based re-identification uses a deterministic per-subject random number generator seeded from the global seed combined with the subject identifier hash (S6.4). U-Net training uses seed 1337 with deterministic CUDA settings (`cudnn.deterministic=True, cudnn.benchmark=False`).

No further seed control is applied to model initialization, `DataLoader` shuffling, augmentation, or batch ordering. Variability across the OASIS-1 5-fold CV and across NIH Chest X-ray bootstrap resamples therefore incorporates both data-partition-level variability and run-level RNG variability inherent to unseeded model initialization and `DataLoader` shuffling. GS transformations are not seeded for individual generation runs (S3.6); the once-per-condition cache fixes the realization used across downstream experiments.

S9.3 Output Mapping

Table S8 maps each main-manuscript figure and table to the experimental section that produces its values, the supplementary section that documents its procedure, and the released-code script or notebook that generates the corresponding output.

S9.4 Compute Profile

Experiment-class wall-clock time on the documented hardware (NVIDIA Tesla P100, single GPU) is approximate: GS dataset generation per (dataset, masking level) takes from minutes to a few hours depending on partition size and follows the runtime characterization in S2.5; utility classifier training takes 1-4 hours per (dataset, architecture, masking level); supervised re-identification classifier training takes 0.5-3 hours per condition; embedding-based re-identification adds negligible additional time on top of utility training because it reuses utility checkpoints; U-Net reconstruction training takes 1-3 hours for dataset and masking level; where DPS fine-tuning and reconstructions are significantly more time consuming (8-12 hours) and computationally costly.

Main Manuscript Output	Experiment (Main)	Supplement Procedure
Table 1 (cross-domain re-id Top-1)	3.1	S3.2.5, S3.3.3, S5, S8.2, S8.4
Table 2 (in-domain re-id Top-1, embedding Rank-1)	3.2	S3.2.5, S3.3.3, S5, S6, S8.2, S8.4
Table 3 (reconstruction SSIM/PSNR/re-id)	3.3	S3.2.6, S3.3.4, S7, S8.2, S8.4, S8.5, S8.6
Table 4 (utility AUC)	3.4	S3.2.4, S3.3.2, S4, S8.2, S8.3
Figure 1 (framework overview)	2.4	S3.5
Figure 2 (cross-domain re-id)	3.1	as Table 1
Figure 3 (in-domain + embedding re-id)	3.2	as Table 2
Figure 4 (reconstruction examples)	3.3	as Table 3
Figure 5 (utility AUC)	3.4	as Table 4
Figure 6 (utility-identity trade-off)	3.5	aggregates Tables 1, 2, 4

Table S8. Mapping from main-manuscript figures and tables to supplementary procedure sections and released-code modules. Module paths are illustrative and reflect the structure of the released repository identified in Section 7 of the main manuscript.