Empirical Privacy Without Differential Privacy: A Privacy–Utility Frontier Analysis of Synthetic CT Generation

Notebook for the CSMorgan Lab at ImageCLEF at CLEF 2026

Md Ragib Shaharear[1,*], Ojonugwa Oluwafemi Ejiga Peter[1], Mahmudul Hoque[1,*], Dapiriye Briggs[2] and Md Mahmudur Rahman[1]

[1]Department of Computer Science, Morgan State University, Baltimore, MD 21251, USA
[2]Electrical and Computer Engineering Department, Morgan State University, Baltimore MD 21251, USA

Abstract

We describe the csmorgan team's participation in all three subtasks of the ImageCLEFmedical GANs 2026 lab, which evaluates privacy leakage and memorisation risks in generative models trained on medical images. Our main positive result comes from **Subtask 3 (Privacy-Preserving CT Slice Generation)**, where a deliberately compact DCGAN-style generator achieved a composite Privacy Preservation Score (PPS) of **0.9086**, placing it among the strongest privacy-preserving submissions in this edition. The same submission obtained a CT-specific FID of 3.24, compared with a 0.32–0.62 range reported for high-utility submissions, locating our model at the privacy-extreme end of the privacy–utility frontier. We attribute this profile to a privacy-oriented training recipe (strong zero-centred R_1 regularisation, a deliberately weak discriminator, EMA weights, GroupNorm, and patience-based early stopping) and we are explicit that, in our submission, this privacy is entangled with reduced fidelity: two of the four official leakage scores are essentially zero (NND audit, white-box inversion) while membership inference and patient re-identification retain non-trivial, above-chance signal. We also report negative but informative results on **Subtask 1 (Detect Training Data Usage)** and **Subtask 2 (Identify Corresponding Latent Vectors)**, and we trace each development-to-test gap to a concrete cause. For Subtask 1, a feature-based ensemble reached an internal ROC AUC of 0.7557, but its decision threshold was selected to maximise Cohen's κ on the same out-of-fold predictions; on the official test phase, all runs achieved near-chance or below-chance κ (best $\kappa = -0.0290$). For Subtask 2, a memory-bank InfoNCE matcher recovered only 2.24% of the 10,000 image–latent matches; we show that the optimistic development accuracy was computed over a pool dominated by training images, whereas the held-out validation curve already predicted the test result.

Keywords

ImageCLEF, ImageCLEFmedical, Generative Adversarial Networks, Privacy-Preserving Synthetic Data, Membership Inference Attack, Latent Space Matching, Privacy–Utility Trade-off

1. Introduction

Synthetic medical imagery from deep generative models is a promising way to share data for AI research without releasing real patient images [1], provided the synthetic distribution does not retain identifiable *fingerprints* of the training data patient anatomy, near-duplicates, or latent representations linking a generated image to a real one. Over multiple editions, the ImageCLEFmedical GANs lab [2, 3, 4] has shown that generative models, both GANs and diffusion-based, can leak such information in practice [5, 6].

The 2026 edition formalises three probes. **Subtask 1** asks whether each real image was used to train a generator observed only through its outputs; **Subtask 2** asks participants to recover, for each of 10,000 synthetic images, the 128-dimensional latent code that produced it from an unordered candidate pool; and **Subtask 3** asks for 5,000 privacy-preserving CT slices, evaluated by four attacks (nearest-neighbour

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21–24, 2026, Jena, Germany
*Corresponding author.
mdsha1@morgan.edu (M. R. Shaharear); ojeji1@morgan.edu (O. O. E. Peter); mahoq1@morgan.edu (M. Hoque); dabri23@morgan.edu (D. Briggs); md.rahman@morgan.edu (M. M. Rahman)
0009-0006-3414-1567 (M. R. Shaharear); 0009-0003-2039-3075 (O. O. E. Peter); 0009-0006-5532-4135 (M. Hoque); 0009-0003-2039-3075 (D. Briggs); 0000-0002-1663-6391 (M. M. Rahman)

audit, membership inference, patient re-identification, white-box inversion). Figure 1 summarises our pipelines for all three.

We describe the csmorgan team's participation in all three subtasks. Our main positive result is Subtask 3, where a compact DCGAN-style generator reached strong empirical privacy under the official metrics; Subtasks 1 and 2 produced negative but informative test results. We therefore frame the paper not as a privacy guarantee, nor as a claim that empirical privacy replaces differential privacy, but as a privacy–utility frontier analysis: strong empirical privacy is achievable under the benchmark attacks, but in our submission it is entangled with a clear loss of CT fidelity. A cross-cutting theme is that several internal estimates were optimistic for an identifiable reason in each subtask, and that telling an honest estimate from an optimistic one is itself a lesson of this benchmark. Our contributions are:

1. For **Subtask 3** (our main result), a deliberately compact, privacy-oriented DCGAN-style generator strong zero-centred R_1 regularisation, EMA weights, GroupNorm, a weak discriminator, and patience-based early stopping achieved a Privacy Preservation Score (PPS) of 0.9086 at a CT-specific FID of 3.2358, at the privacy-extreme end of the frontier. We decompose the composite and show it is carried by two of the four attacks, while membership inference and patient re-identification retain above-chance leakage.
2. For **Subtask 1**, a feature-based ensemble (spectral, statistical, deep ResNet-distance, and shadow-autoencoder cues) reached internal ROC AUC 0.7557 but tuned its threshold to maximise Cohen's κ on the same out-of-fold predictions; all official runs reached near- or below-chance κ (best -0.0290). We argue this, with features measured as absolute distances to a fixed synthetic pool, explains the development-to-test gap.
3. For **Subtask 2**, an end-to-end memory-bank InfoNCE matcher (ConvNeXt-Base encoder, Sinkhorn, CSLS, mutual-kNN, TTA, Hungarian) recovered $224/10{,}000$ matches (2.24%). We show the optimistic development accuracy was measured over a training-dominated pool and used for model selection, while the held-out curve already anticipated the official result.
4. We synthesise the failure modes into one lesson internal estimates leaked information about the predicted quantity in each case and outline a privacy–utility sweep that would separate our Subtask 3 mechanisms from the effect of underfitting, addressing the organisers' open questions.

2. Related Work

Shadow-model membership inference [7] was later extended to generative models [8, 9], on the premise that training samples occupy higher-density or more reconstructable regions than held-out samples. In diffusion models, memorisation can be highly local, with attack success depending strongly on training-set size, capacity, and sampling [10, 11]. The ImageCLEFmedical GANs benchmark brings this question to medical images [5, 6]. Because such attacks hinge on the comparison representation low-level and frequency descriptors capture global artefacts while deep and self-supervised features such as DINOv2 [12] capture anatomical similarity our Subtask 1 pipeline combines spectral, statistical, deep-distance, and reconstruction-error features rather than raw pixels.

Matching a generated image to its latent code probes the generator: if it is smooth and approximately injective, a learned inverse can recover the latent. Contrastive methods (CPC, SimCLR, MoCo) [13, 14, 15] provide the alignment framework, and our Subtask 2 matcher uses a memory-bank InfoNCE objective; as our results show, development alignment need not transfer under distribution shift, non-injectivity, or optimistic evaluation. The dominant formal route to privacy is differential privacy, e.g. via DP-SGD [16], which guarantees privacy but can degrade sample quality under tight budgets. Empirical alternatives R_1 regularisation [17], EMA weights [18], capacity control, and early stopping — give no formal guarantee but can reduce benchmark leakage by discouraging sharp instance-specific fitting; batch normalisation has further been argued to act as an implicit leakage channel [19], motivating batch-statistic-free choices. Our Subtask 3 system explores this regime. Finally, prior work by our group and within the ImageCLEFmedical lineage has used text-guided synthesis, fine-tuned Stable Diffusion, DreamBooth, and LoRA to generate synthetic medical images and training data for colonoscopy and gastrointestinal

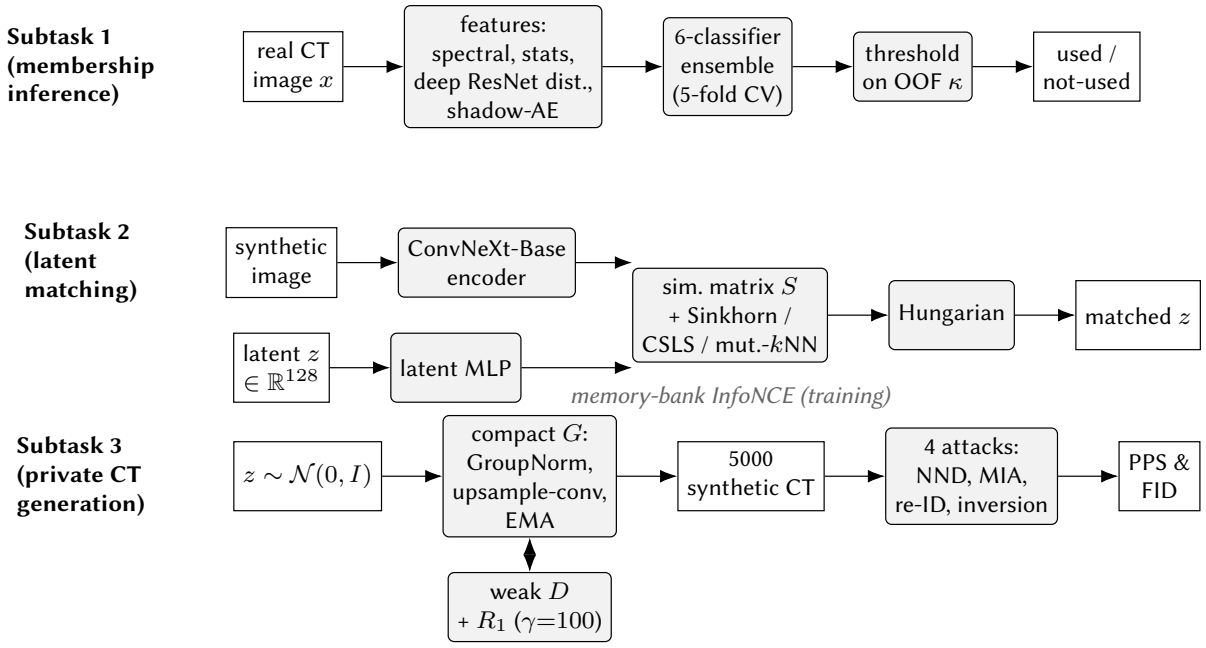

Figure 1: Overview of the csmorgan pipelines for the three subtasks. Subtask 1 stacks complementary features into a classifier ensemble whose threshold is tuned on out-of-fold κ. Subtask 2 embeds images and latents into a shared space (trained with a memory-bank InfoNCE objective) and solves a one-to-one assignment after similarity post-processing. Subtask 3 trains a deliberately compact, strongly regularised generator and submits its samples and checkpoint to four official privacy attacks.

tasks [20, 21, 22, 23]; whereas those target clinically useful images, the present benchmark asks whether generated images preserve training-data fingerprints, whether latents can be linked back, and whether CT synthesis can be made empirically privacy-preserving.

3. Data and Tasks

The 2026 ImageCLEFmedical GANs dataset [2] comprises axial slices extracted from 3D CT scans of approximately 8,000 lung tuberculosis patients. The images are provided as 256×256 8-bit grayscale PNG slices. The challenge is organised into three complementary subtasks that evaluate different aspects of privacy leakage and memorisation in synthetic medical image generation. For **Subtask 1**, participants are given a folder of synthetic images and two folders of real images, `real_used` and `real_not_used`, for training a binary classifier. The goal is to predict whether each image in the held-out `real_unknown` folder was used during the training of the generator. For **Subtask 2**, participants receive 10,000 synthetic images together with an unordered pool of 10,000 candidate 128-dimensional latent vectors. The goal is to output a permutation that matches each generated image to its corresponding latent vector. For **Subtask 3**, participants generate 5,000 synthetic CT slices and submit them together with the generator checkpoint. The submitted images and model are then evaluated using four privacy attacks: nearest-neighbour audit, membership inference, patient re-identification, and white-box inversion [**?**]. All four attacks use a held-out patient set of CT slices never released to participants as a negative control. We report and discuss these official privacy-attack results in Section 5.

4. Methodology

4.1. Subtask 1: Detecting Training Data Usage

Subtask 1 was treated as a binary membership-inference problem: each real CT image had to be classified as either used or not used during generator training. We used a stacked feature-based

ensemble. For each image, we computed spectral, statistical, deep-feature, and reconstruction-error features, standardised them, and trained classical classifiers under a 5-fold stratified cross-validation scheme. First, we computed the azimuthally averaged log power spectrum of each real image x,

$$s(x) = \log\bigl(1 + |\widehat{x}|^2\bigr)\,, \tag{1}$$

and compared it with the mean and standard deviation of spectra from the synthetic image pool G:

$$\mu_G = \mathbb{E}_{g\in G}\left[s(g)\right], \qquad \sigma_G = \mathrm{std}_{g\in G}\left[s(g)\right]. \tag{2}$$

From these spectra we extracted ℓ_1 and ℓ_2 distances, cosine similarity, and standardised z-score statistics. We also included low-level image descriptors such as intensity mean, standard deviation, skewness, kurtosis, percentiles, interquartile range, histogram entropy, and Sobel edge-magnitude statistics. To capture higher-level similarity, we extracted multi-scale features from the conv-3, conv-4, and conv-5 stages of an ImageNet-pretrained ResNet-50 [24]. Each CT slice was resized to 256×256 and replicated across three channels. For each real image, we computed cosine and ℓ_2 distances to its $k = 1, 3, 5, 10, 20$ nearest neighbours in the synthetic image pool. Finally, we reused cached shadow-autoencoder reconstruction-error features from an earlier V3 pipeline. The final feature vector was passed to an ensemble of six classical classifiers: two `HistGradientBoosting` variants (whose histogram-based design follows LightGBM [25]), a `GradientBoostingClassifier`, an `ExtraTreesClassifier` [26], a `RandomForestClassifier`, and an L_2-regularised logistic-regression model [27]. The ensemble was trained using 5-fold `StratifiedKFold` cross-validation, and per-classifier predictions were combined with weights proportional to their out-of-fold (OOF) AUC. A single decision threshold was then selected on the OOF predictions by maximising Cohen's κ. As we discuss in Section 6.4, this choice is consequential: because the official leaderboard metric is itself Cohen's κ, tuning the threshold to maximise OOF κ produces an optimistically biased internal κ even though the threshold-free AUC is unaffected. Most of our features are also *absolute* distances or z-scores measured against a fixed synthetic pool rather than per-image self-normalised quantities, which makes a fixed threshold fragile to any global shift between the development and hidden test distributions.

4.2. Subtask 2: Image–Latent Matching

Subtask 2 was formulated as a $10{,}000$-way matching problem. Given synthetic images and an unordered pool of 128-dimensional latent vectors, the goal was to recover the exact latent vector that generated each image. We learned a shared embedding space for images and latents and then solved a similarity-based assignment problem. The image branch used a ConvNeXt-Base backbone [28], pretrained on ImageNet-22K and fine-tuned on ImageNet-1K, modified for single-channel CT input. A two-layer MLP projected image features into a 256-dimensional embedding space. The latent branch used a three-layer MLP to map each $z \in \mathbb{R}^{128}$ into the same space. Both branches produced ℓ_2-normalised embeddings. Training used a memory-bank InfoNCE objective. For an image embedding u_i, true latent index y_i, and latent memory bank $M \in \mathbb{R}^{N\times D}$ with $N = 10{,}000$, the loss was

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\langle u_i, M_{y_i}\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle u_i, M_j\rangle/\tau)}, \tag{3}$$

with temperature $\tau = 0.05$. The memory bank was updated using a MoCo-style momentum rule [15]:

$$M_{y_i} \leftarrow \mathrm{normalise}\left(0.999\, M_{y_i} + 0.001\, v_i\right), \tag{4}$$

where v_i is the current latent-branch embedding. We trained with AdamW using learning rates of 2×10^{-5} for the ConvNeXt backbone and 5×10^{-4} for the projection heads, weight decay 10^{-4}, 3 warm-up epochs, 40 total epochs, and gradient-norm clipping at 1.0. We split the development set $90/10$ into a training and a held-out validation partition. Checkpoint selection used Top-100 retrieval accuracy on the held-out partition. We emphasise a distinction that proved important for interpreting our results:

the held-out validation partition is an honest generalisation estimate, whereas a separate "development accuracy" that we computed over the full 10,000-image development pool is optimistic, because roughly 90% of that pool consists of images the encoder was trained on. As described in Section 6.5, post-processing methods and their hyperparameters were selected on this full-pool accuracy, which inflated expectations relative to the held-out signal. At inference time, all test images and candidate latents were embedded and compared using the full cosine-similarity matrix S. We optionally applied Sinkhorn normalisation [29], CSLS reweighting [30],

$$\mathrm{CSLS}_{ij} = 2S_{ij} - \bar{S}_i^{\mathrm{row}} - \bar{S}_j^{\mathrm{col}}, \tag{5}$$

mutual-kNN boosting, and multi-crop test-time augmentation. The final one-to-one assignment was obtained with the Hungarian algorithm [31].

4.3. Subtask 3: Privacy-Preserving CT Generation

Subtask 3 required generating $5{,}000$ synthetic CT slices and submitting the generator checkpoint so that the organisers could evaluate both image quality and privacy leakage. Our design prioritised empirical privacy under the benchmark attacks rather than maximising visual fidelity. We used a compact DCGAN-style architecture. The generator projected $z \in \mathbb{R}^{128}$ into a $4 \times 4 \times 8c$ feature map with base channel width $c = 48$, followed by six nearest-neighbour upsample-plus-convolution blocks until reaching 256×256. Each block used two 3×3 convolutions, GroupNorm [32], and LeakyReLU activation. We used GroupNorm instead of BatchNorm because it avoids dependence on batch-level statistics, which has been argued to act as an implicit leakage channel [19]. The final layer used a $\tanh$ projection to produce a single-channel image. The discriminator was a weaker five-stage convolutional downsampler with base channel width 24 and no normalisation. This asymmetric design was intended to reduce sharp discriminator responses around individual training samples while still allowing the generator to learn the global CT structure. Both networks were trained with the non-saturating logistic GAN loss [33]. We applied the zero-centred R_1 gradient penalty [17]

$$\mathcal{R}_1(D) = \frac{1}{2}\mathbb{E}_{x \sim p_{\mathrm{data}}}\left[\|\nabla_x D(x)\|_2^2\right], \tag{6}$$

with $\gamma = 100$, computed every $r = 8$ generator steps and scaled by r to preserve the time-averaged penalty strength. This regularisation discourages sharp discriminator gradients around real training samples, but it does not provide a formal differential-privacy guarantee. We maintained an exponential moving average of the generator weights,

$$\theta_{t+1}^{\mathrm{EMA}} = \rho\, \theta_t^{\mathrm{EMA}} + (1-\rho)\, \theta_{t+1}, \tag{7}$$

with $\rho = 0.999$ [18], and generated all submitted samples using G^{EMA}.

Training used Adam with $\beta = (0.5, 0.999)$, $\mathrm{lr}_G = 10^{-4}$, $\mathrm{lr}_D = 10^{-5}$, batch size 64, and a maximum of 300 epochs. Early stopping monitored a moving average of the *generator* loss, with patience 30 and a minimum training floor of 80 epochs. We note a methodological caveat here: in a non-saturating GAN, the generator loss measures the current adversarial balance rather than sample fidelity, so this criterion does not select for image quality, and we did not compute FID during training for model selection. Combined with the deliberately weak discriminator, this means our checkpoint corresponds to an adversarial-balance plateau rather than a fidelity optimum a point we return to in Section 6.6. At submission time, we sampled $5{,}000$ vectors $z \sim \mathcal{N}(0, I)$, generated 256×256 grayscale PNGs, and submitted them with the generator checkpoint and a minimal `model.py`/`load_model.py` interface for the organisers' white-box inversion attack.

Table 1
csmorgan Subtask 1 results on the official test phase. All five runs were generated by the same V4 pipeline with different ensemble seeds and threshold settings.

Run	Cohen κ	MCC	Acc.	Sens.	Spec.	F1
ID 469	-0.0804	-0.0916	0.4598	0.2196	0.7000	0.2891
ID 468	-0.0570	-0.0612	0.4715	0.2897	0.6533	0.3541
ID 467	-0.0290	-0.0613	0.4855	0.0449	0.9262	0.0802
ID 466	-0.0626	-0.0713	0.4687	0.2299	0.7075	0.3020
ID 465	-0.0804	-0.0916	0.4598	0.2196	0.7000	0.2891

5. Results

5.1. Subtask 1: Detecting Training Data Usage

Table 1 reports our five official test-phase submissions for Subtask 1. The official leaderboard metric is Cohen's κ. All five runs achieved near-chance or below-chance performance, with Cohen's κ ranging from -0.0804 to -0.0290 and accuracy ranging from 0.4598 to 0.4855. The best run by Cohen's κ was ID 467 ($\kappa = -0.0290$), while the best run by F1 was ID 468 ($F1 = 0.3541$). Overall, these results indicate that the proposed membership-inference pipeline did not generalise successfully to the official challenge test set.

To understand why the official result was poor, we also examined the internal validation behaviour of the V4 ensemble pipeline. This internal analysis is shown in Fig. 2. The left plot shows the confusion matrix and the right plot shows the ROC curve, with an internal validation AUC of 0.7557.

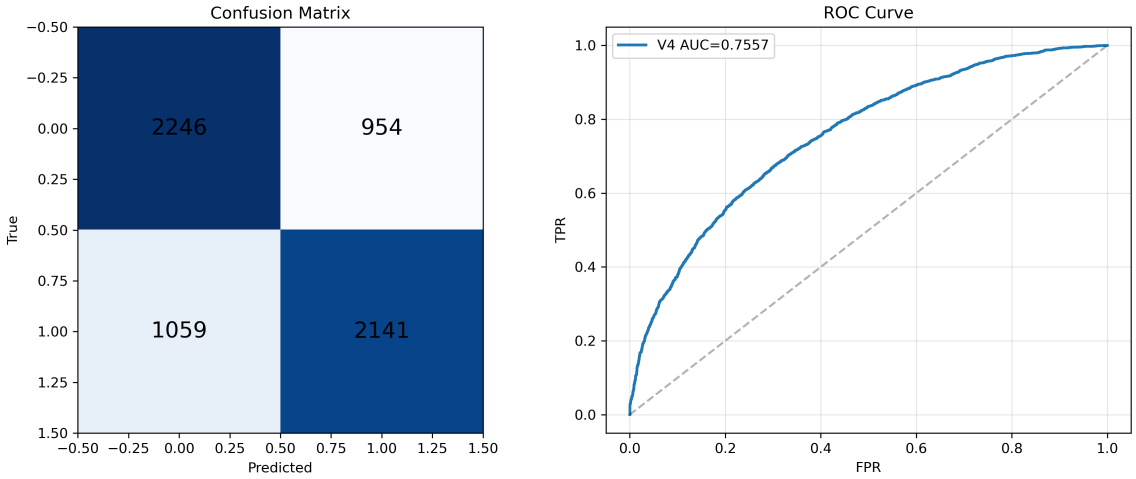

Figure 2: Subtask 1 internal validation results using the V4 pipeline. The left plot shows the confusion matrix, while the right plot shows the ROC curve with an AUC score of 0.7557. These results are from internal validation and should not be interpreted as official test performance.

In the internal validation split, the confusion matrix contained 2246 correct predictions for class 0 and 2141 correct predictions for class 1, with 954 class 0 samples incorrectly predicted as class 1 and 1059 class 1 samples incorrectly predicted as class 0. The internal ROC AUC of 0.7557 indicates some development-set separability. Crucially, however, AUC is threshold-free, whereas the official metric (κ) is not: our threshold was selected to maximise OOF κ on the same predictions, so the favourable internal κ was optimistically biased by construction. We therefore interpret Fig. 2 as evidence of internal separability rather than challenge-set generalisation, and we analyse the development-to-test gap in detail in Section 6.4.

Table 2
csmorgan Subtask 2 results on the official test phase, selected runs.

Run	Matching accuracy	Correct	Wrong
task2_v8_multibackbone	0.0003	3	9997
task2_v11_hybridv2	0.0048	48	9952
task2_v10_e2e_membank	0.0094	94	9906
task2_v10_e2e_membank2	0.0094	94	9906
vversion10raw	0.0107	107	9893
version10	0.0108	108	9892
wo_stage_matcher_m4max	0.0224	224	9776

5.2. Subtask 2: Image–Latent Matching

Table 2 reports a representative subset of our official Subtask 2 test-phase submissions. The leaderboard metric is strict matching accuracy: a prediction is counted as correct only if the image is assigned to its exact generating latent vector among 10,000 candidates.

The best official submission, `wo_stage_matcher_m4max`, recovered 224 out of 10,000 correct image–latent matches, corresponding to 2.24% accuracy. A uniformly random permutation would recover approximately 1 correct match in expectation, or an accuracy of $1/10{,}000 = 0.0001$. Thus, the best submission was above random chance, but still far from reliable exact latent recovery. During development, some configurations of the V10 pipeline produced much stronger retrieval behaviour under an optimistic evaluation protocol that we characterise precisely in Section 6.5. The honest signal came from the held-out validation partition shown in Fig. 3. The training loss decreased steadily from approximately 9.0 to 1.4 over 40 epochs, indicating that the model learned the training pairs, while the validation loss remained nearly flat and slightly increased after the early epochs, indicating overfitting to the training image–latent correspondences.

The held-out Top-K curves support this interpretation, and in retrospect they predicted the official outcome. Top-1 accuracy remained below 1% throughout training consistent with the 2.24% official result — while Top-10 accuracy reached approximately 3–4% before saturating. Coarse recall was more encouraging: Top-50 recall reached approximately 9–10% and Top-100 recall approximately 14–17%. This indicates that the embedding space captured some global alignment between images and latent vectors, but not enough fine-grained information for reliable exact 10,000-way matching. The mean-rank curve also improved early and then degraded, consistent with overfitting after the middle of training. Figure 4 shows an additional diagnostic evaluation of the learned image–latent embedding space. This analysis is useful for understanding the model's internal ranking behaviour, but it should not be read as official test performance. In this diagnostic setting, correct pairs had a higher average similarity score than incorrect pairs, and many true latents were ranked near the top of the candidate list. However, the stronger ranking behaviour observed in this diagnostic analysis did not translate into robust official test matching.

Overall, Subtask 2 shows that the learned representation contained some image–latent alignment signal, but the signal was not strong or stable enough to solve the exact matching problem under the official test conditions.

5.3. Subtask 3: Privacy Attacks

Table 3 reports the four official privacy-attack leakage scores and the composite Privacy Preservation Score (PPS) for our Subtask 3 submission. This was the strongest result among our three subtasks.

The four leakage scores are not uniform, and we read the composite through its components. The nearest-neighbour distance (NND) audit produced a ratio of 0.9787 (mean synthetic-to-training distance 17.73 versus synthetic-to-held-out 18.12 in InceptionV3 feature space), giving leakage $\ell_1 = 0.0213$: synthetic images were essentially equidistant from their nearest training and held-out neighbours. The 162 images flagged by the per-image memorisation threshold represent less than 4% of the submission

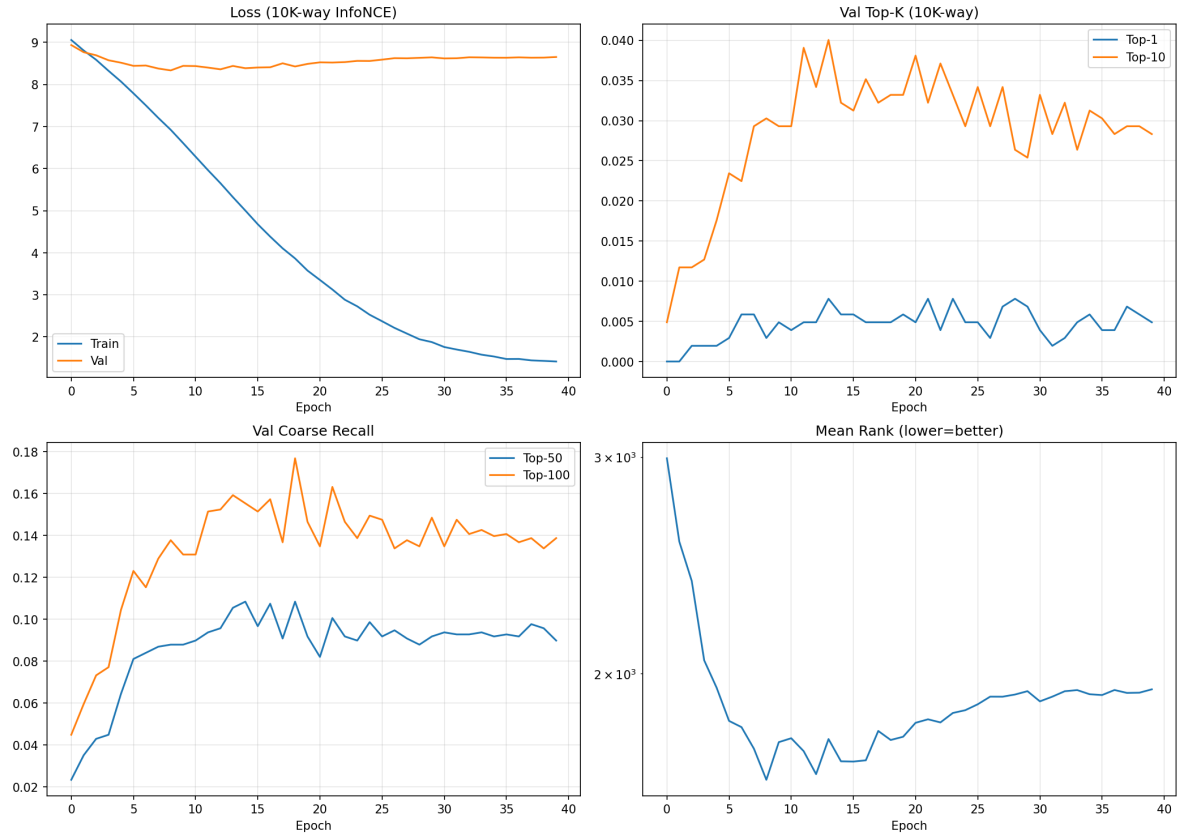

Figure 3: Subtask 2 training and validation behaviour under the held-out validation setting. The figure shows the 10K-way InfoNCE loss, validation Top-K accuracy, coarse recall, and mean rank across training epochs.

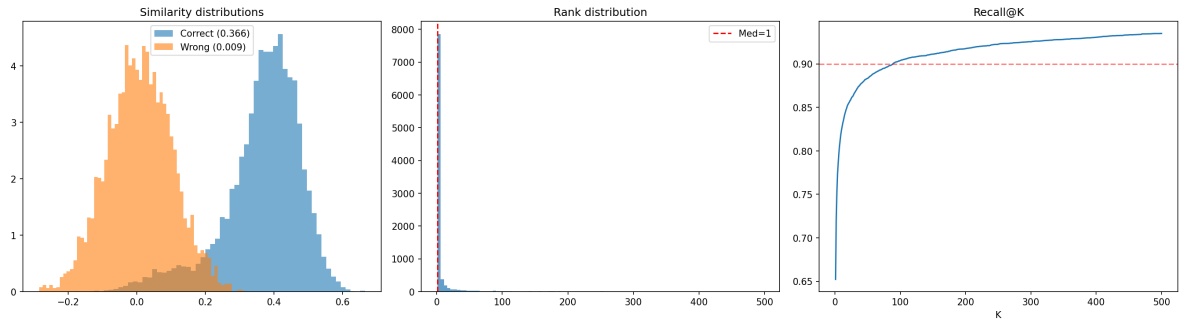

Figure 4: Subtask 2 diagnostic evaluation of the image–latent matching model. The left plot shows similarity distributions for correct and wrong pairs, the middle plot shows the rank distribution of the correct latent vectors, and the right plot shows Recall@K. This diagnostic analysis is not the official test result.

and may partly reflect natural overlap between training and held-out CT distributions rather than direct instance copying. The white-box inversion attack likewise found no signal: mean reconstruction error was 17.82 for training images and 17.74 for held-out images, a difference of only 0.43%, yielding a slightly negative held-out-minus-train delta of -0.0762 and $\ell_5 = 0.0000$.

The remaining two attacks are different. The membership-inference attack produced an AUC-ROC of 0.5916 and TPR@5%FPR of 8.6% ($\ell_2 = 0.1832$): above a perfectly random operating point, though far from strong leakage. More notably, the patient re-identification attack produced rank-1 coverage of 0.1611 and rank-5 coverage of 0.3020, against a chance level of $1/P = 0.0067$ for $P = 149$ patients, with a normalised assignment entropy of 0.4185 (a uniform assignment would be near 1.0). Re-identification coverage of about $24\times$ chance is therefore reduced but clearly not eliminated. As the organisers note, Attack 3 is precisely the leakage that output-level filtering cannot remove, because

Table 3
csmorgan Subtask 3 official privacy results. PPS is the composite $1 - \mathrm{mean}(\ell_1, \ell_2, \ell_3, \ell_5)$ as defined by the organisers; their attack numbering skips ℓ_4, so the white-box inversion leakage score is denoted ℓ_5. Per-attack metrics are reported as released by the task organisers. Patient re-identification is over $P = 149$ patients, giving a chance level of $1/P = 0.0067$.

Attack	Headline metric	Value	Leakage ℓ	Interpretation
A1 NND audit	NND ratio R	0.9787	0.0213	equidistant
A2 Membership inf.	AUC-ROC	0.5916	0.1832	above chance
A3 Patient re-ID	rank-1 coverage	0.1611	0.1611	$\approx 24\times$ chance
A4 White-box inversion	Δ (held-out$-$train)	-0.0762	0.0000	no signal
FID (CT evaluator)	$\downarrow$		3.2358	
PPS	$\uparrow$		**0.9086** strong privacy	

non-trivial coverage indicates that patient-specific anatomy was internalised during training. The honest reading of our PPS is thus that the composite is carried almost entirely by ℓ_1 and ℓ_5 (near zero), while ℓ_2 and ℓ_3 retain measurable, above-chance signal. Even a privacy-extreme, heavily smoothed generator does not fully erase distributional or patient-identity information only the instance-copy and latent-inversion signals.

Together, these four attacks produced a PPS of 0.9086, which places the submission in the organisers' "strong privacy" band ($\mathrm{PPS} \geq 0.80$) and among the most privacy-preserving submissions in this edition. However, this came with a clear utility cost: a CT-specific FID of 3.2358, substantially higher than the 0.32–0.62 range reported for high-utility submissions, locating our model at the privacy-extreme end of the privacy–utility frontier rather than at the high-fidelity end. Figure 5 shows representative synthetic CT slices from the submitted generator. The images preserve the broad structure of axial chest CT slices, including the body boundary, lung regions, heart region, spine, ribs, and soft-tissue regions, while showing visible smoothing and reduced fine-detail sharpness. This visual behaviour is consistent with the quantitative results: the generator captures global CT structure while suppressing highly specific patient-level detail, which improves the instance-level privacy metrics while degrading fidelity.

Figure 6 shows the training behaviour of the Subtask 3 generator. The discriminator and generator losses remained stable, with the discriminator loss around 1.38 and the generator loss around 0.69, suggesting that the adversarial training process did not collapse. The R_1 penalty was initially higher and then decreased gradually, indicating that the discriminator gradients around real samples became smoother during training, consistent with the intended design.

The right plot in Fig. 6 shows training time per epoch. Most epochs required a relatively consistent amount of time, with a small number of spikes likely caused by checkpoint saving or system-level scheduling overhead; because the loss curves remained stable, the timing variation did not appear to affect learning.

6. Discussion: the Privacy–Utility Trade-off

The most informative outcome of our participation is not a uniformly strong performance across all three subtasks, but the contrast between them. Subtask 3 produced a strong empirical privacy result under the official benchmark attacks, whereas Subtasks 1 and 2 produced negative but informative official test results. A second, cross-cutting theme is that in each subtask our most optimistic internal number was optimistic for an identifiable reason: the estimate was computed in a way that let information about the quantity being predicted leak into the estimate itself. We make this precise below.

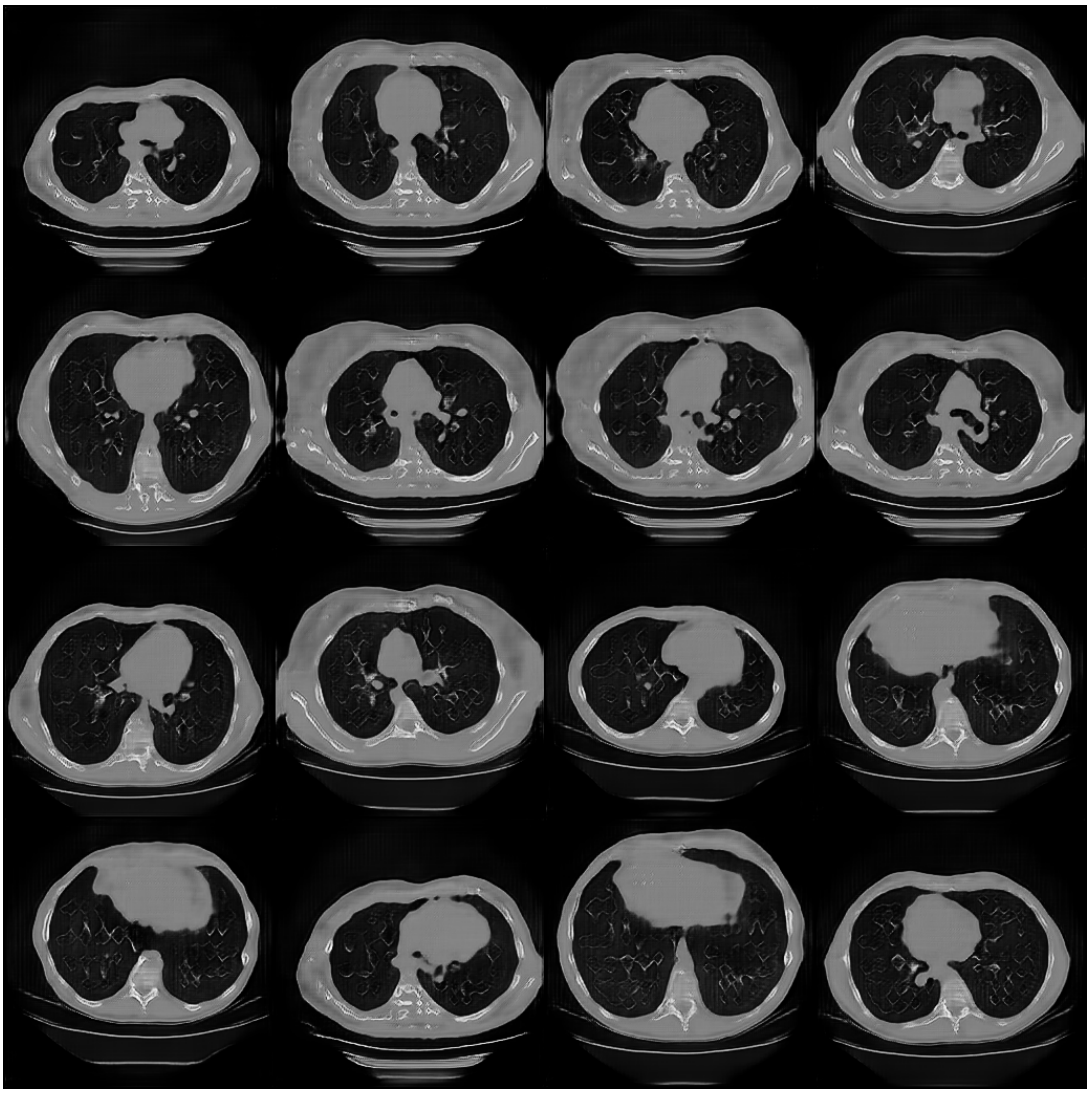

Figure 5: Synthetic CT slices generated by the Subtask 3 privacy-oriented generator. The generated samples preserve the general axial lung CT structure, including body boundary, lung regions, heart region, and internal thoracic anatomy, while showing visible smoothing and reduced fine-detail sharpness due to the privacy-oriented training strategy.

6.1. Subtask 3 as a privacy-extreme operating point

Our Subtask 3 submission achieved a PPS of 0.9086, in the organisers' "strong privacy" band. As decomposed in Section 5.3, this composite is driven by two near-zero leakage scores (NND audit and white-box inversion) and offset by two non-trivial ones (membership inference at $\ell_2 = 0.1832$ and patient re-identification at $\ell_3 = 0.1611$, the latter about $24\times$ chance). The strong empirical privacy was accompanied by a CT-specific FID of 3.2358, substantially higher than the 0.32–0.62 range reported for high-utility submissions. Our model should therefore not be read as a high-fidelity synthetic CT generator; it is a privacy-extreme point on the frontier whose samples preserve broad axial CT structure but lose fine anatomical detail.

6.2. Design choices contributing to the privacy–utility profile

We caution that the following attributions are hypotheses about a single operating point, not conclusions from a controlled comparison; the sweep we propose in Section 6.6 is what would test them. The discriminator was intentionally weaker than the generator, using half the base channel width and a $10\times$ lower learning rate, which reduced the strength of the adversarial signal and likely discouraged the

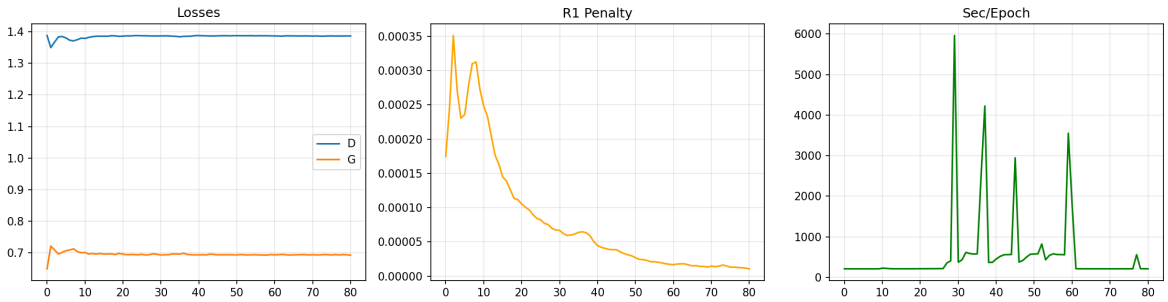

Figure 6: Training behaviour of the Subtask 3 generator. The left plot shows the discriminator and generator losses, the middle plot shows the R_1 regularisation penalty, and the right plot shows the training time per epoch.

generator from fitting highly specific training examples at the cost of a less precise signal for reproducing fine anatomical structure, contributing to the higher FID and visible smoothing. The zero-centred R_1 penalty with $\gamma = 100$ encouraged smoother discriminator gradients around real samples, which may help explain the low instance-level leakage, but overly strong regularisation can also flatten useful distribution cues and limit fine texture. EMA weights and patience-based early stopping stabilised the generator but may also have prevented it from reaching the fidelity of utility-focused submissions. We used GroupNorm rather than BatchNorm to avoid batch-statistic dependence, which has been argued to act as an implicit leakage channel [19]; we do not claim this choice alone guarantees privacy.

6.3. Interpreting the negative inversion delta

The white-box inversion delta is one of the more interesting Subtask 3 results, but it is best read as the absence of signal rather than as privacy in the held-out direction. The mean reconstruction errors were 17.82 (training) and 17.74 (held-out); the held-out-minus-train delta of -0.0762 is under half a percent of the absolute error scale and is most naturally interpreted as noise around zero. A plausible explanation is that the generator learned a smoothed CT-like manifold that is not especially close to either set at the instance level, so the inversion attack is shaped more by the global geometry of the learned manifold than by training membership. This is consistent with the visual results.

6.4. Why Subtask 1 did not generalise

Subtask 1 highlights the difficulty of image-level membership inference against synthetic medical images, but our specific failure has a more concrete diagnosis than "calibration drift" alone. The threshold-free internal AUC was 0.7557, yet the official metric, Cohen's κ, collapsed to near or below chance. The decisive factor is that we selected the decision threshold to maximise OOF κ on the very predictions used to report it; since κ is also the leaderboard metric, the favourable internal κ was optimistically biased, while the AUC being threshold-free was not. This is consistent with internal separability that fails to translate into a transferable operating point. A second factor is that most of our features are absolute distances or z-scores against a fixed synthetic pool rather than per-image self-normalised statistics, so a fixed threshold is fragile to any global intensity or preprocessing shift between the development data and the hidden `real_unknown` set, which were provided as separate releases. A third, and more sobering, possibility is that an internal AUC of 0.7557 that vanishes on the official set partly reflects an incidental difference between the `real_used` and `real_not_used` folders (for example in acquisition or preprocessing) rather than true membership signal; such a confound would be absent in `real_unknown`. A useful diagnostic for future editions is to test whether the two training folders are separable using only features that should be membership-irrelevant (e.g. a plain intensity histogram with no reference to the synthetic pool); separability there would indicate a folder artefact rather than membership leakage. For these reasons we interpret the internal ROC result as evidence of development-set separability only.

6.5. Why Subtask 2 did not transfer

Subtask 2 also revealed a large development–test gap, and here too we can name the mechanism. Our optimistic "development accuracy" including the strongest Top-100 recall figures from earlier pipeline versions was computed over the full $10{,}000$-image development pool, of which roughly 90% were images the encoder had been trained on, since we trained on a 90% partition and evaluated retrieval over the whole pool. Worse, the choice of post-processing method (raw, CSLS, Sinkhorn, mutual-kNN, or their combinations) and its hyperparameters was made by maximising accuracy on that same contaminated pool and then transferred to the test set. The held-out validation partition, by contrast, was an honest estimate: its Top-1 accuracy stayed below 1% throughout training and correctly anticipated the 2.24% official result. We therefore attribute the gap primarily to training-image contamination of the development metric and to model/post-processing selection on it, rather than to distribution shift alone. Two additional factors plausibly contribute and are not mutually exclusive: non-injectivity of the generator, whereby visually similar slices map to different latents and frustrate exact recovery even when coarse alignment exists; and genuine generator-distribution differences between the development and test releases. The diagnostic curves show that some global alignment existed; the signal was simply too weak and too dependent on the development protocol to solve exact $10{,}000$-way matching.

6.6. From one operating point to a frontier

A limitation we wish to state plainly is that our Subtask 3 evidence is a single operating point, and that the strong privacy and the low fidelity are confounded: the same blurring that raises FID to 3.24 is also the most parsimonious explanation for the near-zero instance-level leakage. Our attributions in Section 6 (to R_1, the weak discriminator, EMA, and GroupNorm individually) cannot be separated from this underfitting on the basis of one run. The task organisers raise exactly this question in their feedback, asking which design choices led to the outcome and whether the privacy properties could be preserved at higher fidelity. We believe the appropriate way to answer is a privacy–utility sweep: varying a single control at a time for example the R_1 weight $\gamma \in \{0, 10, 100\}$, the discriminator width, or the early-stopping patience to produce several $(\mathrm{PPS}, \mathrm{FID})$ points, and adding a matched-fidelity baseline (a vanilla DCGAN at comparable FID) to test whether our regularisation reduces leakage *beyond* what blur alone provides. Because each run is the same small generator with one changed hyperparameter, this is inexpensive, and it would convert "frontier analysis" from a single dot into an actual frontier. We did not run this sweep in time for submission and report it as the primary direction for follow-up work; the organisers' forthcoming overview paper, which proposes a joint privacy–utility evaluation framework, is the natural venue against which to position such results.

6.7. Implications for empirical privacy without differential privacy

Our results should be interpreted carefully. They do not show that empirical privacy mechanisms can replace formal differential privacy: differential privacy provides a mathematical guarantee, whereas our approach provides only benchmark-level empirical evidence under a specific set of attacks, and even that evidence is uneven across the four attacks. What the results do show is that architectural capacity control, regularisation, EMA inference, and early stopping can move a generator toward a low-leakage operating point under the ImageCLEFmedical GANs 2026 evaluation at a substantial fidelity cost, and without eliminating distributional or patient-identity leakage. The generated images may be useful as a privacy-stress-test data point at one extreme of the spectrum, but they are not suitable for downstream clinical AI training without substantial fidelity improvement. In practical synthetic medical data deployment, privacy and utility must be evaluated jointly rather than optimising one metric in isolation.

7. Conclusion

We described the csmorgan team's participation in all three subtasks of the ImageCLEFmedical GANs 2026 lab. The main positive result came from **Subtask 3**, where a compact, privacy-oriented DCGAN-style generator achieved a Privacy Preservation Score of 0.9086 under the official benchmark attacks at a CT-specific FID of 3.2358, placing it at the privacy-extreme end of the privacy–utility frontier. We were explicit that this composite is carried by two near-zero leakage scores while membership inference and patient re-identification retain above-chance signal, and that the strong privacy is, on a single operating point, confounded with underfitting. For **Subtask 1**, a feature-based ensemble reached an internal ROC AUC of 0.7557 but did not transfer: all official runs achieved near-chance or below-chance Cohen's κ (best $\kappa = -0.0290$), which we trace to a threshold tuned to the leaderboard metric on out-of-fold data and to absolute-distance features sensitive to development–test shift. For **Subtask 2**, the InfoNCE memory-bank matcher recovered only 224 out of the correct matches $10{,}000$ (2.24%); we showed that the optimistic development accuracy was measured in a training-dominated pool, while the hold-out validation curve already predicted the official result. The common lesson across all three subtasks and our main methodological contributionis that an internal estimate is only as honest as its isolation from the quantity it predicts, and we outline a privacy–utility sweep that would let future work separate genuine privacy mechanisms from the effect of reduced fidelity.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) to brainstorm pipeline designs, refine code, and draft and polish text. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgments

This work used computational resources provided by the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program under Allocation Request **CIS251242**, including **53.0 GPU hours on the NCSA Delta GPU system**. The author acknowledges the National Science Foundation (NSF), ACCESS, and NCSA Delta for supporting the computational resources used in this research.

8. Online Resources

GitHub Repository: https://github.com/Ejigsonpeter/Empirical-Privacy-Without-Differential-Privacy

References

[1] Yao, Zhaomin, Zhen Wang, Ying Zhan, Xiaodan Wu, Wei Zhang, Yusong Pei, Guoxu Zhang, and Zhiguo Wang. "Applications of generative models in medical imaging: A review." *Engineering Applications of Artificial Intelligence*, 2026.

[2] B. Ionescu, H. Müller, D.-C. Stanciu, A. Radzhabov, A. García Seco de Herrera, A.-G. Andrei, A. Băicoianu, et al., "ImageCLEF 2026: Multimodal Challenges in Medicine, Science, Agritech, and Security," in: *Advances in Information Retrieval*, Proceedings of the 48th European Conference on Information Retrieval, ECIR 2026, Lecture Notes in Computer Science, vol. 16486, Springer Nature Switzerland, Cham, 2026, pp. 336–344.

[3] A.-G. Andrei, M.-G. Constantin, M. Dogariu, D. Karpenka, Y. Prokopchuk, A. Radzhabov, D.-C. Stanciu, L.-D. Ştefan, V. Kovalev, H. Müller, and B. Ionescu, "Overview of the 2026 ImageCLEFmedical

GANs task: Fingerprint detection, latent space analysis, and privacy-preserving medical image synthesis," in: *CLEF 2026 Working Notes*, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, Sept. 21–24, 2026.

[4] A.-G. Andrei, A. Radzhabov, I. Coman, V. Kovalev, B. Ionescu, and H. Müller, "Overview of ImageCLEFmedical GANs 2023 Task – Identifying Training Data 'Fingerprints' in Synthetic Biomedical Images Generated by GANs for Medical Image Security," *CLEF 2023 Working Notes*, CEUR Workshop Proceedings, CEUR-WS.org, vol. 3497, pp. 1305–1315, 2023.

[5] A.-G. Andrei, A. Radzhabov, D. Karpenka, Y. Prokopchuk, V. Kovalev, B. Ionescu, and H. Müller, "Overview of the 2024 ImageCLEFmedical GANs Task – Investigating Generative Models' Impact on Biomedical Synthetic Images," *CLEF 2024 Working Notes*, CEUR Workshop Proceedings, CEUR-WS.org, vol. 3740, pp. 1412–1428, 2024.

[6] A.-G. Andrei, M.-G. Constantin, M. Dogariu, A. Radzhabov, L.-D. Ştefan, Y. Prokopchuk, V. Kovalev, H. Müller, and B. Ionescu, "Overview of ImageCLEFMedical 2025 GANs Task: Training Data Analysis and Fingerprint Detection," in: *CLEF 2025 Working Notes*, CEUR Workshop Proceedings, CEUR-WS.org, vol. 4038, pp. 2182–2196, 2025.

[7] R. Shokri, M. Stronati, C. Song, V. Shmatikov, "Membership inference attacks against machine learning models," in: *IEEE Symposium on Security and Privacy*, 2017, pp. 3–18.

[8] J. Hayes, L. Melis, G. Danezis, E. De Cristofaro, "LOGAN: Membership inference attacks against generative models," *Proc. on Privacy Enhancing Technologies*, 2019(1):133–152.

[9] D. Chen, N. Yu, Y. Zhang, M. Fritz, "GAN-Leaks: A taxonomy of membership inference attacks against generative models," *CCS*, 2020.

[10] N. Carlini, et al., "Extracting training data from diffusion models," *USENIX Security*, 2023.

[11] J. Duan, et al., "Are diffusion models vulnerable to membership inference attacks?," *ICML*, 2023.

[12] M. Oquab, et al., "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[13] A. van den Oord, Y. Li, O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[14] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, "A simple framework for contrastive learning of visual representations," *ICML*, 2020.

[15] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, "Momentum contrast for unsupervised visual representation learning," *CVPR*, 2020.

[16] M. Abadi, et al., "Deep learning with differential privacy," *CCS*, 2016, pp. 308–318.

[17] L. Mescheder, A. Geiger, S. Nowozin, "Which training methods for GANs do actually converge?," *ICML*, 2018.

[18] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, "Analyzing and improving the image quality of StyleGAN," *CVPR*, 2020.

[19] Wu, Yuxin, and Justin Johnson. "Rethinking" batch" in batchnorm," *arXiv preprint*, 2021.

[20] O. O. Ejiga Peter, M. M. Rahman, and F. Khalifa, "Advancing AI-Powered Medical Image Synthesis: Insights from the MedVQA-GI Challenge Using CLIP, Fine-Tuned Stable Diffusion, and DreamBooth + LoRA," *arXiv preprint arXiv:2502.20667*, 2025.

[21] O. O. Ejiga Peter, M. Hoque, F. A. Ejiga, and R. N. Chowdhury, "Solving Medical Data Limitations Through AI: Multi-Modal Vision-Language Learning for Gastrointestinal VQA and Synthetic Training Data Generation," in: *CLEF 2025 Working Notes*, CEUR-WS.org, 2025.

[22] O. O. Ejiga Peter, O. T. Adeniran, J.-O. A. MacGregor, F. Khalifa, and M. M. Rahman, "Text-Guided Synthesis in Medical Multimedia Retrieval: A Framework for Enhanced Colonoscopy Image Classification and Segmentation," *Algorithms*, 18(3):155, 2025. doi:10.3390/a18030155.

[23] O. O. Ejiga Peter, B. D. Tunde, M. Hoque, D. Briggs, M. M. Rahman, and F. Khalifa, "Colonoscopy Image Synthesis: Transforming Medical Image Classification With Enhanced Quality and Computational Efficiency," in: *2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI)*, 2025, pp. 1–6. doi:10.1109/ICMI65310.2025.11141284.

[24] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," *CVPR*, 2016.

[25] G. Ke, et al., "LightGBM: A highly efficient gradient boosting decision tree," *NeurIPS*, 2017.

[26] P. Geurts, D. Ernst, L. Wehenkel, "Extremely randomized trees," *Machine Learning*, 63(1):3–42, 2006.
[27] F. Pedregosa, et al., "Scikit-learn: Machine learning in Python," *JMLR*, 12:2825–2830, 2011.
[28] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, "A convnet for the 2020s," *CVPR*, 2022.
[29] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *NeurIPS*, 2013.
[30] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, H. Jégou, "Word translation without parallel data," *ICLR*, 2018.
[31] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.
[32] Y. Wu, K. He, "Group normalization," *ECCV*, 2018.
[33] I. Goodfellow, et al., "Generative adversarial networks" *NeurIPS*, 2014.