

Analyzing CycleGAN Training Behavior with Offline Average Loss for CBCT-to-CT Image Translation

Mahya Ahmadzadeh¹ · Keyur Shah² · Gregory C. Sharp, PhD³ · Nagarajan Kandasamy, PhD¹ · John MacLaren Walsh, PhD¹

¹Electrical and Computer Engineering Department, Drexel University, Philadelphia, PA

² Emory University, Atlanta, GA

³ Massachusetts General Hospital, Boston, MA

ABSTRACT

When has a CycleGAN actually converged?

For unpaired CBCT-to-CT translation, the per-batch adversarial loss is too noisy to read. The batch size one and Keras' moving average couple each report to the optimizer's trajectory, not the model's true quality. We introduce an offline average loss: a decoupled read-out that freezes a checkpoint, disables gradients, and scores the full dataset for one clean value per epoch, yielding a smooth, monotone convergence curve. We use this metric to compare training schedules, and to fairly judge which trained model is better: train a fresh discriminator on each generator and evaluate it only inside the patient anatomy. Together, these tools turn CycleGAN training from a guessing game into a measurable, principled process.

INTRODUCTION

Problem: CBCT is used daily in radiation therapy but has poor contrast, making it unsuitable for dose recalculation. CycleGAN can synthesize CT-quality images, but its training loss is unreliable.
Limitation: Standard fixes like early stopping introduce bias, skewing model comparisons.
Solution: We propose an offline average loss for tracking convergence and a constrained fresh discriminator for fair model evaluation.

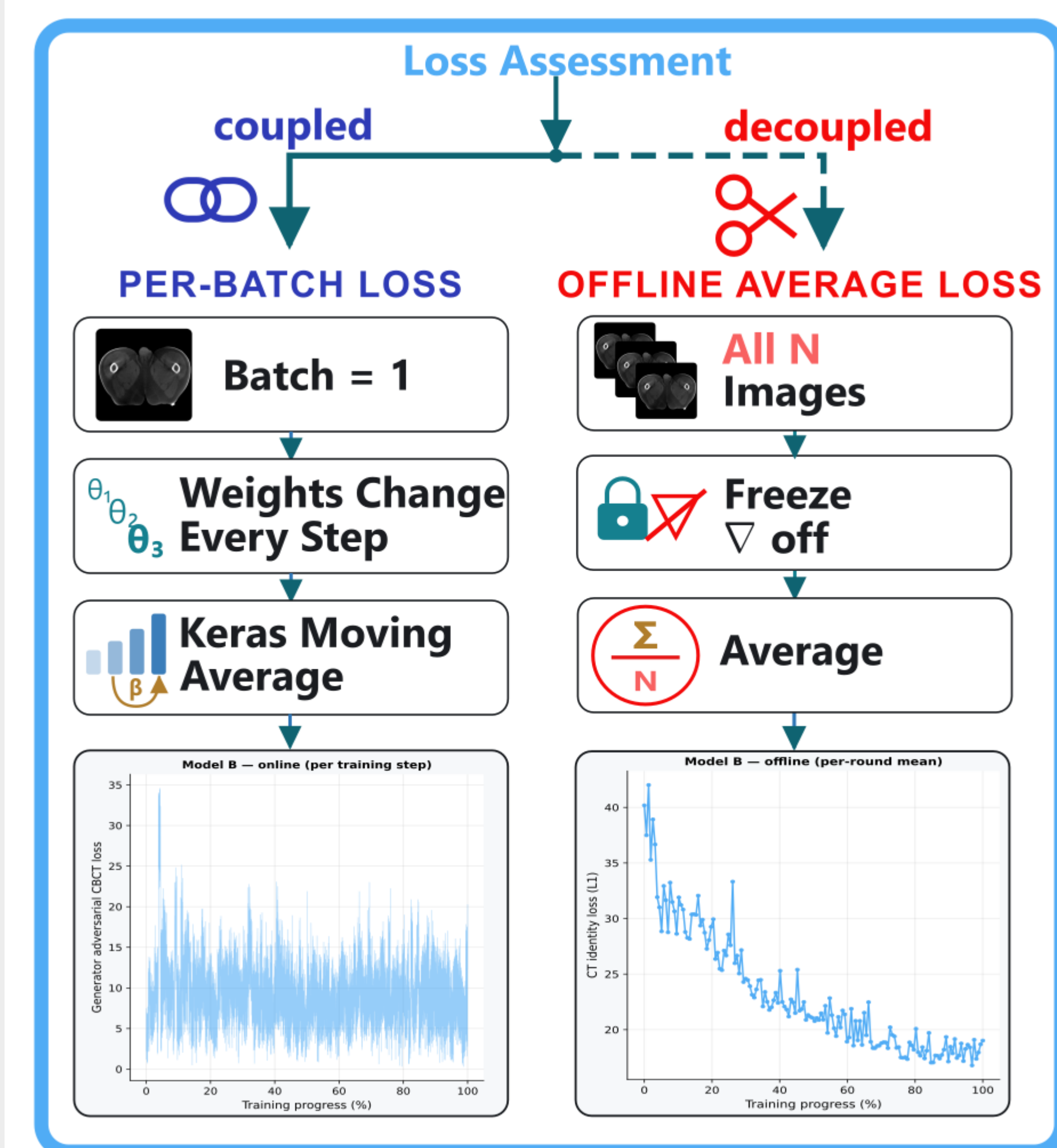


Figure 1: Raw per-step losses (left) vs. one-epoch averaged losses (right).

METHODS AND MATERIALS

Data: 58 pelvic cancer patients • 72 slices • 256×256.

Split: 32 train / 9 val / 17 test (random seed=42).

Two models compared:

Model A: patched weighted-mask training, 50 epochs

Model B: aligned CBCT-CT pairs, 158 rounds

Offline average loss: Computed by freezing checkpoints and disabling gradients to find the average CT-identity loss (L1) over the full dataset. This provides one clean value per epoch without affecting training.

Constrained fresh discriminator: For each generator checkpoint, a fresh PatchGAN is initialized and trained for exactly 30 epochs (no early stopping). AUC and P_{MD} are evaluated at $P_{FA} = 5\%$ and 10% on 17 held-out patients, averaged across 3 seeds (mean ± 1 std).

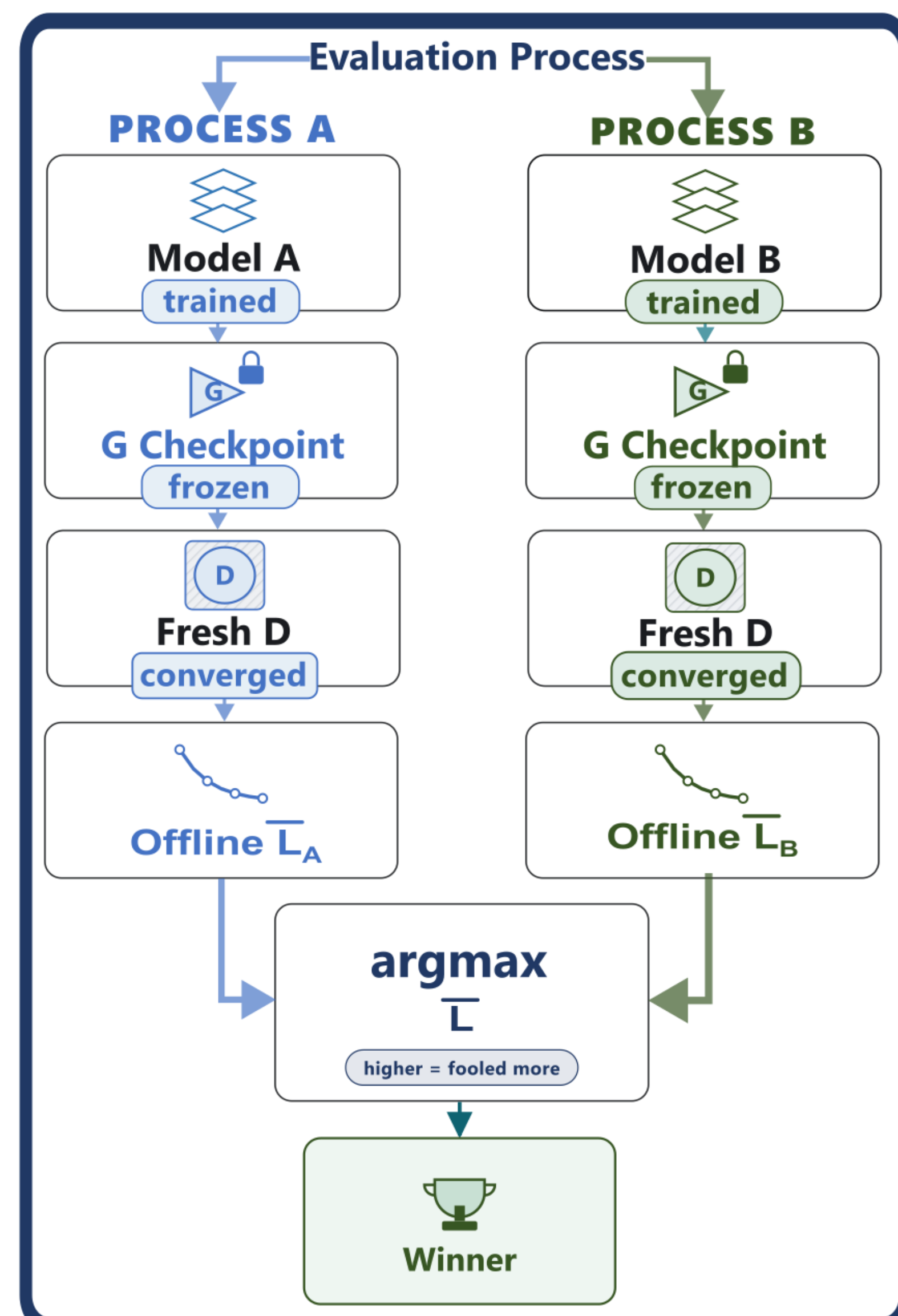
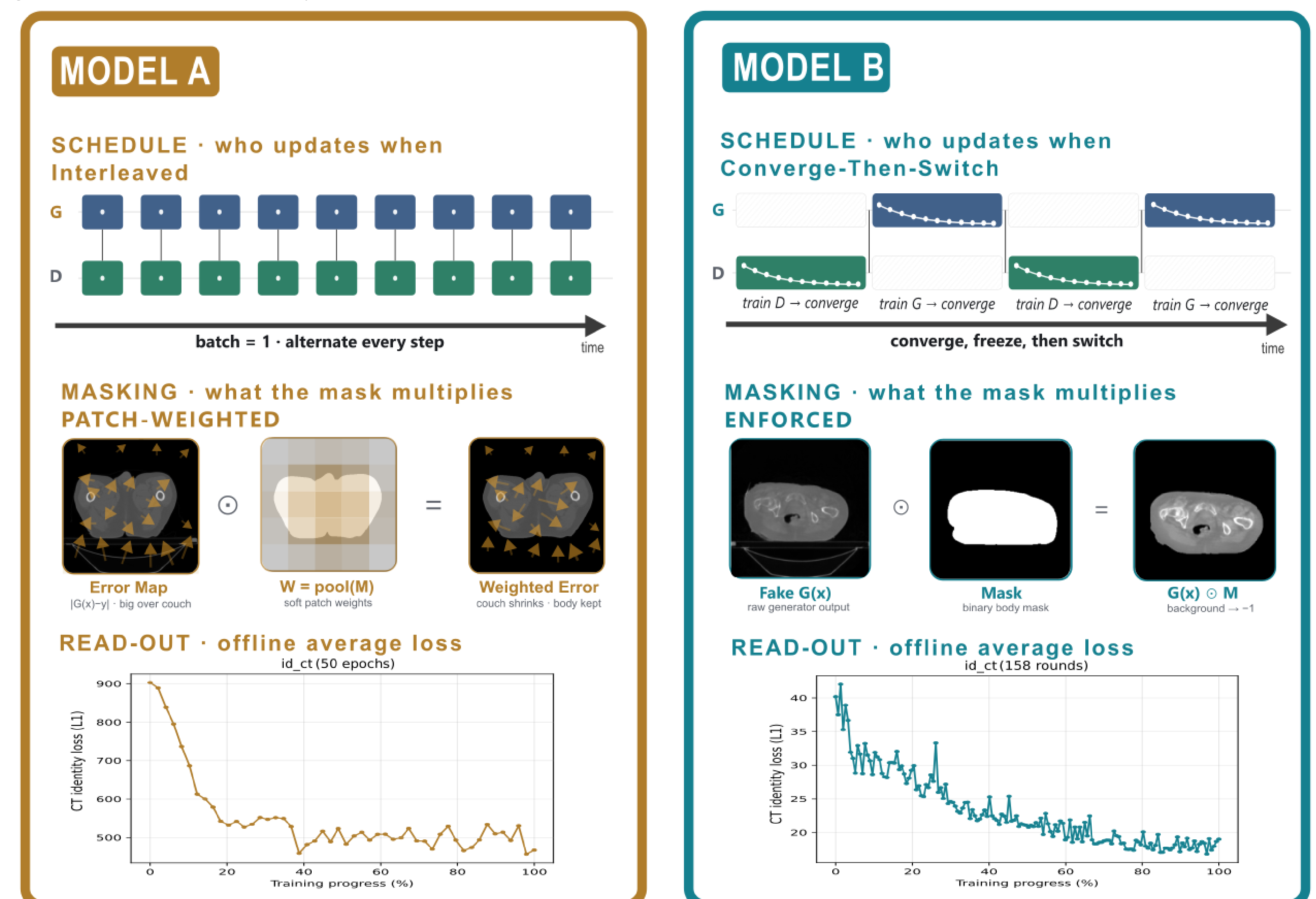


Figure 2: Evaluation workflow: training a fresh PatchGAN discriminator for 30 epochs per generator checkpoint.

Discussion

To fairly evaluate different training strategies, we implement an identical evaluation protocol across two distinct training regimes, comparing their schedule, masking, and offline loss metrics side by side.



Results

Early Training: A 5-epoch discriminator easily separates fakes from both models ($AUC = 1$).

Final Checkpoint $d = 10$:

Model A: Remains perfectly detectable ($AUC = 1.000$).

Model B: Significantly more realistic; AUC drops to 0.765 ± 0.027 and roughly two-thirds of fakes escape detection ($P_{MD} = 0.681 \pm 0.021$ at $P_{FA} = 5\%$).

Budget Impact: A tight budget $d = 5$ is fooled by both models, whereas a loose budget $d = 30$ successfully detects both.

Takeaway & Stability: The crossover at $d = 10$ proves Model B requires far more discriminator effort to detect, indicating higher realism. Findings are stable across 3 random seeds (mean ± 1 std).

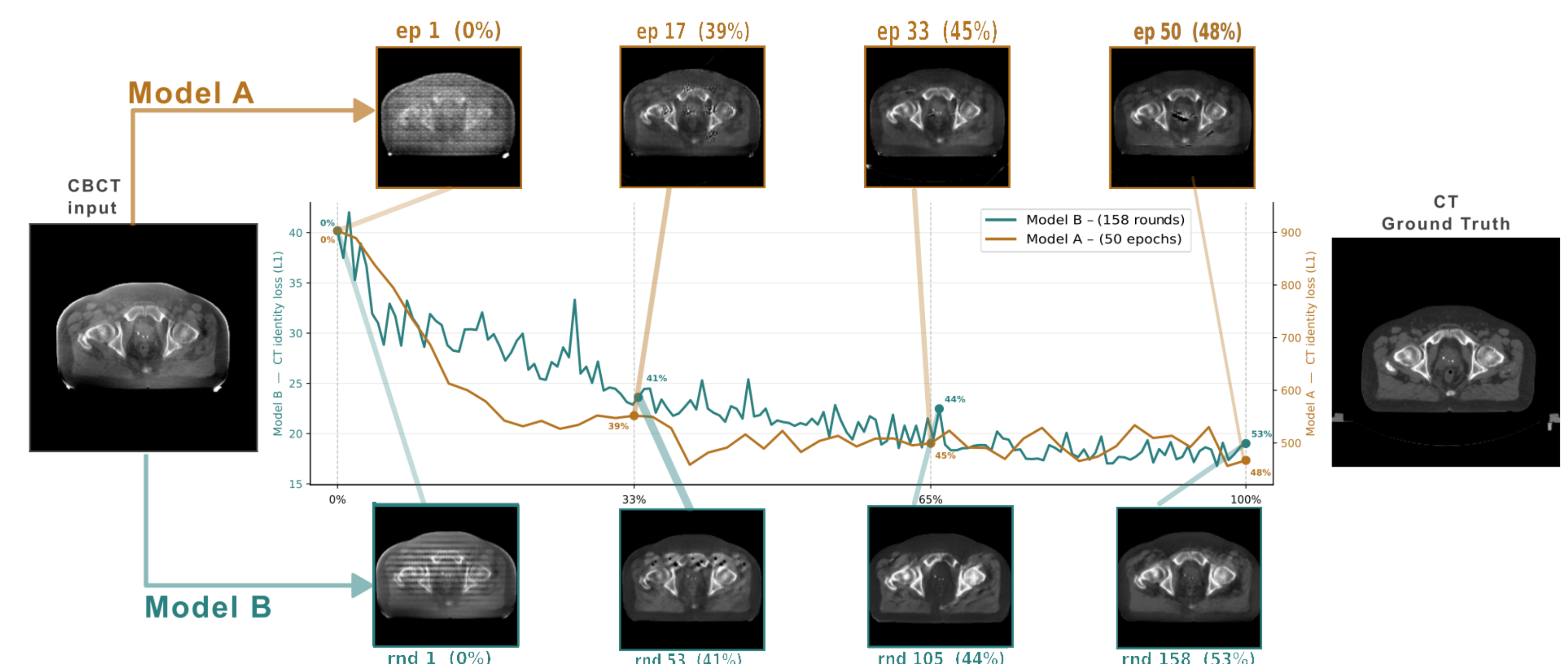


Figure 3: Realism comparison. Model B is significantly harder to detect than Model A at budget $d = 10$.

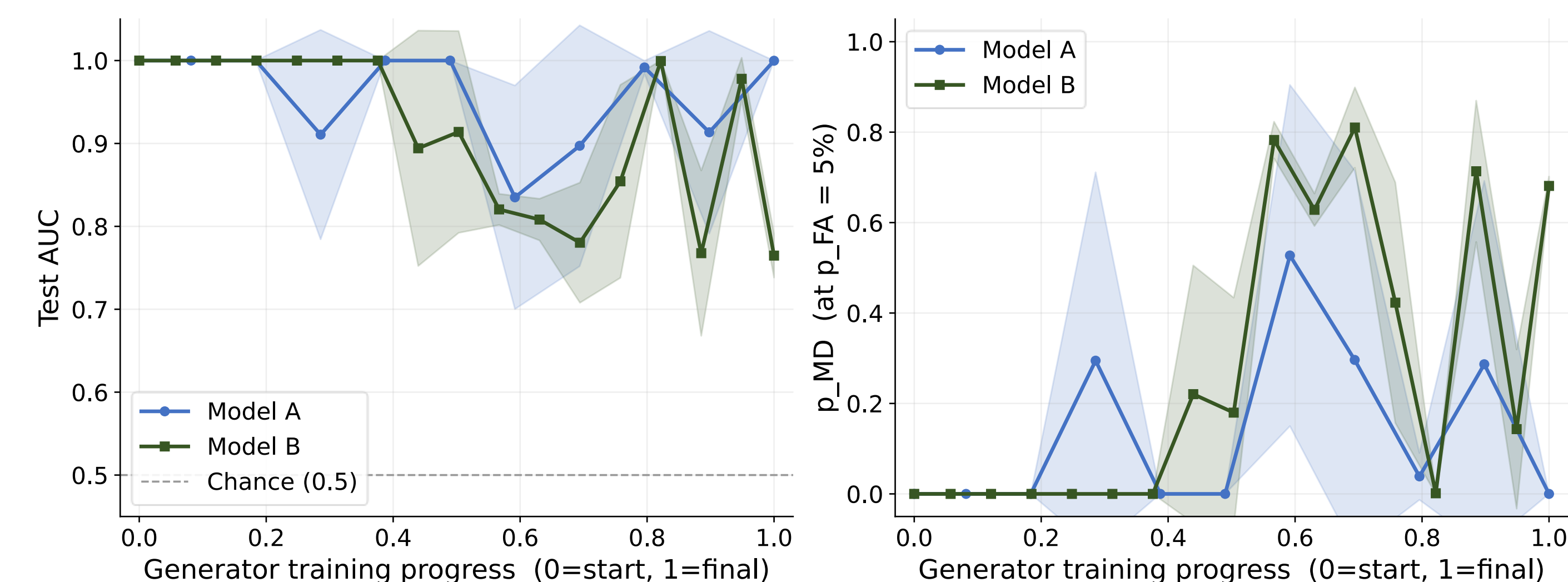


Figure 4: Trained fresh discriminator (budget = 10 disc epochs, $n_{train} = 32$ and $n_{test} = 17$ and mean ± 1 std over 3 seeds to find the winner generator.

CONTACT

Mahya Ahmadzadeh

Drexel University

Philadelphia, USA

ma3923@drexel.edu

267-228-2347