Imagined Memorisation: Training-Data Leakage in Model-Based RL World Models

Anonymous Authors[1]

Abstract

Model-based reinforcement learning (MBRL) agents such as DreamerV3 (Hafner et al., 2025) and IRIS (Micheli et al., 2023) train a *world model*—a pixel-generative sequence model—on heavily revisited replay buffers, a regime that maximises memorisation risk. We present the first systematic membership-inference audit of MBRL world models, adapting three attack families (trajectory reconstruction, dynamics-loss MIA, and adversarial-action divergence) to the action-conditioned generative setting and evaluating across DreamerV3 and IRIS on four Atari games. On IRIS / Ms. Pac-Man, reconstruction-based MIA attains AUC$=0.999$ with Cohen's $d = -4.76$ and TPR$=0.98$ at 1% FPR—exceeding signals typically reported for language and diffusion models—yet standard loss-based MIA flags zero members on the same checkpoint, and five of eight loss-MIA evaluations score below random. We attribute this disagreement to a collection-policy state-space mismatch that swamps likelihood-based scores while leaving pixel-level signals intact. The implication is that memorisation in pixel-generative world models concentrates in the decoder pathway—the inverse of the language-model setting in which loss-based MIA is the standard tool.

1. Introduction

A world model $p_\theta(o_{t+1}, r_{t+1} \mid o_{\leq t}, a_{\leq t})$ is the central learned component in modern MBRL. In contrast to policy networks, world models are high-capacity generative sequence models trained on *small, heavily revisited* replay buffers—the Atari 100k replay is $\sim$400k frames consumed for $\sim$600 epochs in IRIS (Micheli et al., 2023). By the duplication scaling laws characterised by Carlini et al. (2022), this regime maximises memorisation risk.

Existing privacy work in reinforcement learning targets either *policy* networks via action-MIA (Gomrokchi et al., 2021) or text-conditioned diffusion models (Chen et al., 2024); the world model itself, the largest learned component in modern MBRL, has not been audited. The gap is consequential: proprietary world models trained on driving footage (GAIA-1/2; Hu et al. 2023), robot demonstrations (RT-2, π_0; Brohan et al. 2023; Black et al. 2024), and gameplay are deployed behind APIs whose outputs are directly accessible to adversaries.

Contributions.

1. We introduce three black-box MIA primitives that operate over any action-conditioned world model's encode/imagine/decode interface (Section 4).
2. We report the first memorisation audit of MBRL world models, covering both recurrent (DreamerV3) and transformer (IRIS) architectures across four Atari games (Section 5).
3. We show that reconstruction-based and likelihood-based MIA disagree systematically on the same checkpoints, and trace the disagreement to a collection-policy distribution-shift confound that pixel-level scores resist (Sections 5, 6).

2. Related Work

Memorisation in generative models. Verbatim memorisation in language models was first characterised via extraction attacks (Carlini et al., 2019) and later quantified at scale by Carlini et al. (2022), who established the TPR-at-low-FPR reporting convention we adopt. Memorisation in diffusion-based generators has been studied in image (Carlini et al., 2023) and video (Chen et al., 2024) settings.

Privacy in reinforcement learning. Existing RL privacy work targets *policy* networks via action-level MIA (Gomrokchi et al., 2021) or studies population-level dataset inference. To our knowledge, the world model itself—the high-capacity sequence generator that dominates training

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the FoGen Workshop at ICML 2026. Do not distribute.

compute in modern MBRL (Micheli et al., 2023; Hafner et al., 2025)—has not been audited under any membership-inference framework prior to this work.

3. Background and Threat Model

World models. IRIS (Micheli et al., 2023) represents trajectories as token sequences produced by a VQ-VAE tokeniser and modelled by a GPT, with each block laid out as $[o_t^{(0)}, \ldots, o_t^{(K-1)}, a_t]$. DreamerV3 (Hafner et al., 2025) couples a convolutional encoder/decoder with an RSSM latent dynamics module over discrete categorical variables. Both architectures are trained with a per-pixel reconstruction objective on replay-buffer data.

Threat model. The adversary is given a trained world model and a candidate trajectory $\tau = (o_{1:T}, a_{1:T})$ and must decide whether τ belongs to the training replay buffer. We assume white-box weight access, but the attacks themselves require only the world model's encode/imagine/decode interface and are therefore architecture-agnostic.

4. Attacks

Attack 1: Trajectory Reconstruction. For a candidate trajectory $\tau = (o_{1:T}, a_{1:T})$ we encode $k{=}5$ context frames, imagine H steps under the *true* subsequent action sequence, and score decoded frames against the held-out continuation in LPIPS (Zhang et al., 2018):

$$s_{\text{recon}}(\tau) = \frac{1}{H} \sum_{t=k+1}^{k+H} \text{LPIPS}\big(\hat{o}_t,\, o_t\big), \tag{1}$$

where $\hat{o}_t$ is the decoded imagination at step t. Members are expected to score systematically lower than non-members. We report AUC and Cohen's d at $H \in \{15, 30, 45\}$.

Attack 2: Dynamics-Loss MIA. We threshold the length-normalised one-step prediction NLL (Carlini et al., 2022),

$$s_{\text{loss}}(\tau) = -\frac{1}{N(\tau)} \sum_{t=1}^{T-1} \log p_\theta\big(o_{t+1} \,\big|\, o_{\le t},\, a_{\le t}\big), \tag{2}$$

where $N(\tau)$ is the per-token count for IRIS or the per-pixel count for DreamerV3; normalisation removes the trajectory-length confound. We report AUC and TPR at a 1% false-positive threshold.

Attack 3: Adversarial-Action Divergence. Following the divergence-from-training framing of Nasr et al. (2023), we drive imagination from a real encoded context with one of three adversarial policies π—constant NOOP, high-entropy random, and low-entropy repeat—and score the resulting rollout $\hat{o}^\pi(\tau)$ by its nearest-neighbor LPIPS to the training set,

$$s^\pi_{\text{div}}(\tau) = \min_{\tau' \in \mathcal{D}_{\text{train}}} \text{LPIPS}\big(\hat{o}^\pi(\tau),\, o'_{1:H}\big). \tag{3}$$

A one-sided Kolmogorov–Smirnov test between member and non-member s^π_{div} distributions tests whether rollouts collapse onto training data more readily than onto held-out data. Initialising from real encoded contexts, rather than random latents, removes a confound in which OOD initialisations dominate the score.

Table 1. Per-game MIA results. Recon AUC at $H{=}45$, Cohen's d at $H{=}30$; Loss TPR at 1% FPR; Div. p is min KS p across adversarial-action policies. Bold: AUC$\ge$0.65, TPR$\ge$0.10, or $p{<}0.01$.

		Recon		Loss-MIA		Div.
Model	Game	AUC	d	AUC	TPR@1%	p
IRIS	Pong	0.534	−0.10	**0.680**	0.091	0.030
IRIS	Breakout	**0.692**	−0.60	0.590	0.059	1.000
IRIS	Krull	0.558	−0.00	0.344†	0.077	0.572
IRIS	Pac-Man	**0.999**	**−4.76**	0.434	0.000	1.000
Dreamer	Pong	0.229†	+0.87*	0.356†	0.000	**1.1e-04**
Dreamer	Breakout	0.540	+0.00*	0.636	0.062	0.071
Dreamer	Krull	**0.682**	−0.51*	0.081†	0.000	**1.6e-10**
Dreamer	Pac-Man	0.574	−0.23*	0.303†	0.040	1.000

†AUC<0.5 indicates that members are scored *worse* than non-members; see Section 6.
*Cohen's d approximated from the H=30 AUC via the normality approximation in Appendix A; per-window LPIPS scores were not retained for the original Dreamer reconstruction sweep.

5. Experiments and Results

Setup. Both IRIS and DreamerV3 were originally developed and benchmarked on Atari; our evaluation operates entirely within their native training domain. We evaluate IRIS and DreamerV3-S, both trained from scratch, on Pong, Breakout, Krull, and Ms. Pac-Man; training from scratch gives ground-truth membership labels over the replay buffer, following standard MIA evaluation practice (Carlini et al., 2022). Members are drawn uniformly from the training replay buffer; non-members are 200 trajectories per game collected by a uniform-random policy under a disjoint seed. We use $k{=}5$ context frames, horizons $H \in \{15, 30, 45\}$, 200 windows per condition, and AlexNet-feature LPIPS throughout. All experiments run on H200 GPUs.

Reporting conventions. We report Cohen's d (pooled-std. standardised mean difference; see Appendix A) alongside AUC and TPR at fixed FPR (Carlini et al., 2022). Where per-window scores were not retained (Dreamer reconstruction), d is approximated from AUC under a normality assumption (Appendix A), reproducing directly-measured IRIS values within ± 0.05.

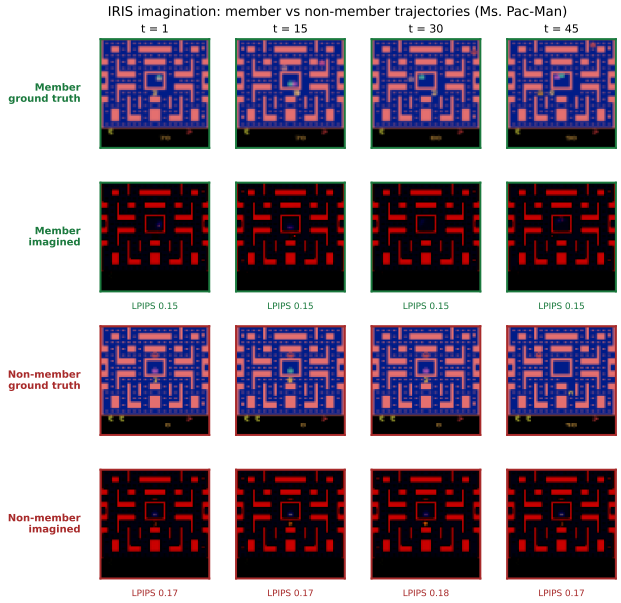

Figure 1. IRIS imagination on Ms. Pac-Man. Top two rows: ground truth and imagined trajectory for a member episode; bottom two rows: same for a non-member. Per-frame LPIPS shown beneath each imagined frame.

Reconstruction memorisation is large where it occurs. On IRIS / Ms. Pac-Man we measure AUC$=$ 0.999 at $H{=}45$ (Cohen's $d = -4.76$; member LPIPS 0.157 vs. non-member 0.169; TPR@1%FPR$=$ 0.98). To calibrate: Carlini et al. (2022) treat $d{>}1$ as a strong memorisation signal in language models, and values above 4 are uncommon in any modality we are aware of. Figure 1 visualises the effect directly—imagined member trajectories track their ground truth at horizons where non-member rollouts have already collapsed structurally.

Cross-attack agreement is the exception, not the rule. None of the eight configurations we evaluate yields a coherent positive signal across all three attack families. The configuration with the strongest cross-attack agreement is DreamerV3 / Krull, on which reconstruction (AUC$=$ 0.682) and adversarial divergence ($p < 10^{-10}$ under every action policy) corroborate membership while loss-MIA on the same checkpoint is severely inverted (AUC$=$ 0.081, TPR@1%FPR$=$ 0.000). The opposite extreme is IRIS / Ms. Pac-Man, where reconstruction reaches AUC$=$ 0.999 while loss-MIA flags zero members at the 1%-FPR threshold (Table 1). With 200 windows per condition, the gap is far too wide to be sampling variance. We treat this disagreement, visualised in Figure 2, as the central empirical finding of this work: reconstruction's AUC$=$ 0.999 confirms that memorisation is present in the checkpoint, so loss-MIA's failure here represents a genuine blind spot rather than a null result.

Loss-MIA inherits a collection-policy confound. Five of the eight configurations register sub-chance loss-MIA AUC—DreamerV3 / Pong, Krull, and Pac-Man together with IRIS / Krull and Pac-Man—and the inversion is extreme in two cases (DreamerV3 / Krull, AUC$=$ 0.081; DreamerV3 / Pong, AUC$=$ 0.356). Sub-chance AUC indicates a reversed score direction: members are scored as *harder* to predict than non-members. We attribute the effect to the collection process. On Pong, per-frame pixel statistics for members and non-members are nearly identical (std 69.6 vs. 69.5), yet mean episode length differs by $3\times$ (1573 vs. 500). A uniform-random policy terminates almost immediately and rarely escapes early-game states; the trained policy reaches diverse mid- and late-game states that random play visits with vanishing probability. Members therefore inhabit a systematically harder-to-predict distribution than non-members, and likelihood-based scores conflate "intrinsically harder to predict" with "not memorised." Perceptual reconstruction metrics largely absorb this confound because LPIPS is comparatively decoupled from state-space difficulty.

Adversarial divergence is teacher-force-dependent. On the IRIS / Ms. Pac-Man configuration that yields recon AUC$=$ 0.999, the divergence attack returns $p \approx 1$ under every adversarial-action policy. Per-window nearest-neighbor LPIPS distributions for member and non-member *contexts* are nearly indistinguishable; once a real context is supplied, adversarial rollouts converge to a common manifold regardless of whether the context originated in the training set. The reconstruction signal is therefore *teacher-forced*—it depends on the true future actions, not on the encoded context alone.

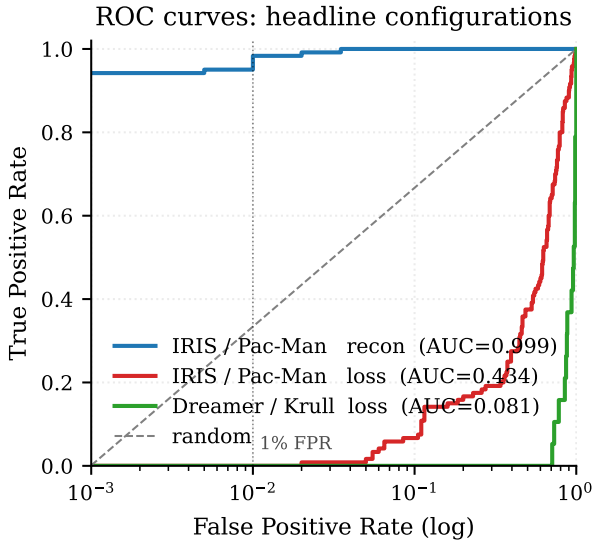

Figure 2. ROC curves for three headline configurations on a log-scale FPR axis. Reconstruction on IRIS / Ms. Pac-Man (blue) reaches TPR$=$ 0.98 at FPR$=$ 1%; loss-MIA on the same checkpoint (red) sits on the diagonal. The dotted line marks the 1%-FPR threshold reported in our table and abstract.

6. Discussion and Limitations

Where does memorisation live? The pattern across our eight configurations is consistent with a single hypothesis: memorisation in pixel-generative world models is concentrated in the decoder pathway—the codebook and convolutional decoder for IRIS, the convolutional reconstruction head for DreamerV3—rather than in the autoregressive likelihood surface. Reconstruction reads the decoder output and registers large effects; loss-MIA reads the likelihood and registers essentially none. This inverts the situation in language models, where the autoregressive likelihood is the canonical memorisation probe.

Effect size in context. Reported d values for language-model memorisation typically fall in $[0.5, 2]$ even for heavily duplicated sequences (Carlini et al., 2022); values exceeding 4 are uncommon across modalities. The $d = -4.76$ we measure on IRIS / Ms. Pac-Man at $H{=}30$ is therefore not merely statistically significant: as Figure 3 shows, the member and non-member LPIPS distributions barely overlap, and the gap is already saturated by $H{=}15$. Together with TPR@1%FPR$= 0.000$ on the same checkpoint's loss-MIA score, this configuration exhibits leakage among the largest documented for any generative model, with its attack surface concentrated in a region that likelihood-based audits do not expose.

A codebook-centred mechanism for the disagreement. The most natural explanation for the recon-vs-loss disagreement, at least in IRIS, is the VQ-VAE codebook. The codebook is a discrete dictionary of visual primitives shared across all training frames—a high-bandwidth, low-frequency channel into which training-specific image patterns can be compiled. The GPT operates on token indices, so its autoregressive likelihood is calibrated over the (small) token vocabulary and is largely indifferent to whether the underlying patches were observed during training. The decoder, by contrast, emits pixel patterns whose fine structure depends on which codebook entries are activated and in what spatial arrangement. Reconstruction reads these fingerprints; loss-MIA does not. We leave a formal architectural decomposition to future work.

Implications for MIA on sequential models. Our results indicate that, for world models trained against an environment, loss-based MIA without policy-matched non-members can produce systematically misleading conclusions. Episode-length and state-coverage gaps between collection policies constitute a distribution shift that interacts with likelihood-based scores even when no adversary is present. For practitioners auditing proprietary world models, we recommend (i) pairing loss-based scores with at least one perceptual-distance score, (ii) constructing non-member sets via collection policies matched to the training distribution—typically shadow training runs with disjoint seeds rather than uniform-random rollouts—and (iii) reporting TPR at low FPR thresholds in addition to AUC, which can mask the tail-regime inversion we observe on Ms. Pac-Man.

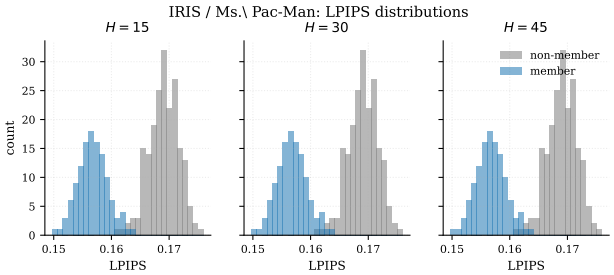

Figure 3. Per-window LPIPS distributions for IRIS / Ms. Pac-Man at three imagination horizons. Member trajectories (blue) cluster around LPIPS≈ 0.157; non-members (grey) around 0.169. The separation is established by $H{=}15$ and stable through $H{=}45$, consistent with memorisation being realised in the decoder rather than accruing across imagination steps.

Limitations. (1) Atari at $64{\times}64$ is a deliberately conservative proxy for real-world threat models; we expect leakage to be *larger* on natural images (driving footage, robot demonstrations) where training data is less compressible. (2) Non-members use a uniform-random policy rather than a policy-matched shadow run. (3) We do not evaluate any defence, including differential-privacy training.

Future work: non-generative world models. Both architectures we audit are pixel-generative: DreamerV3 reconstructs observations through a convolutional decoder, and IRIS reconstructs them through a VQ-VAE. If our central finding—decoder-pathway memorisation—is causal, then *non-generative* world models, which predict latent representations of future observations rather than pixels, should exhibit qualitatively different leakage profiles. JEPA-style world models (LeCun, 2022; Assran et al., 2023) train an encoder to predict the embedding of a future frame given the current context, never reconstructing pixels. Under the decoder-locus hypothesis, JEPA models should be substantially harder to attack with our reconstruction primitive while remaining susceptible to dynamics-loss attacks operating in the latent space. A direct comparison would either corroborate the decoder-locus claim or reveal that the leakage is intrinsic to the dynamics module—we view this as the most informative single follow-up.

References

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Carlini, N., Liu, C., Erlingsson, U., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, 2019.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *USENIX Security Symposium*, 2023.

Chen, L. et al. Investigating memorization in video diffusion models. *arXiv preprint arXiv:2410.21669*, 2024.

Gomrokchi, M., Amin, S., Aboutalebi, H., Wong, A., and Precup, D. Membership inference attacks against temporally correlated data in deep reinforcement learning. *arXiv preprint arXiv:2109.03975*, 2021.

Hafner, D., Lillicrap, T., Norouzi, M., and Ba, J. Mastering diverse domains through world models. *Nature*, 2025.

Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Sheridan, A., Shotton, J., and Kendall, A. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

LeCun, Y. A path towards autonomous machine intelligence. *OpenReview preprint*, 2022. Version 0.9.2.

Micheli, V., Alonso, E., and Fleuret, F. Transformers are sample-efficient world models. In *International Conference on Learning Representations*, 2023.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

A. Metric Definitions

Cohen's d. For score distributions $\{s_M\}$ over members and $\{s_N\}$ over non-members, the standardised mean difference is

$$d = \frac{\bar{s}_M - \bar{s}_N}{\sqrt{(\hat{\sigma}_M^2 + \hat{\sigma}_N^2)/2}}. \tag{4}$$

Negative d indicates members score lower than non-members (expected for reconstruction, where lower LPIPS is better).

AUC-to-d approximation. Where per-window scores are unavailable, d is approximated under a normality assumption (Carlini et al., 2022):

$$d \approx \sqrt{2}\,\Phi^{-1}\big(\mathrm{AUC}\big), \tag{5}$$

where Φ^{-1} is the standard-normal quantile function. Applied to IRIS rows where both estimates are available, this approximation reproduces directly-measured values within ± 0.05.