

Temporal Gradient Inversion for Private Trajectory Reconstruction in Embodied Reinforcement Learning

Sudip Bhujel¹ Shanghao Shi² Ruiquan Huang¹ Ning Zhang² Yang Xiao¹

¹University of Kentucky

²Washington University in St. Louis

NeurIPS 2026



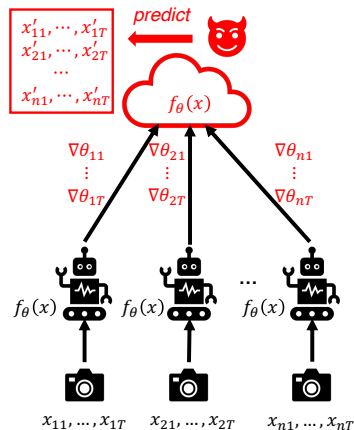
trace-rl.github.io



github.com/trace-rl/TRACE

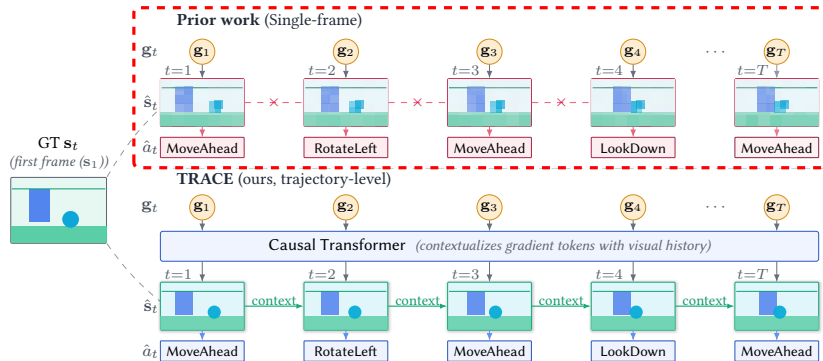
Embodied Agents Share Gradients, Not Observations

- ▶ Robot fleets keep camera frames on-device and send only policy gradients
- ▶ Gradient inversion recovers inputs from gradients
- ▶ A leaked trajectory reveals room layout, objects, and where the agent went



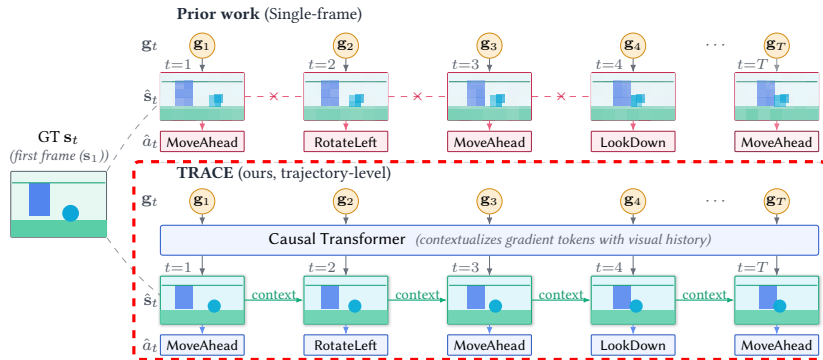
A curious server sees each client's ordered gradients

Prior Attacks Invert Each Gradient Independently



- ▶ Each gradient $\mathbf{g}_t$ is inverted on its own, as an i.i.d. sample
- ▶ Ignores that consecutive frames share layout, objects, and lighting

TRACE Inverts the Whole Gradient Stream



- ▶ Encodes every gradient and attends over the trajectory so far
- ▶ Decodes the frame $\hat{s}_t$ and the action $\hat{a}_t$ at each step

Threat Model: An Honest-but-Curious Server

Server knows

- ▶ Architecture, broadcast checkpoint θ_0 , action space, learning rule
- ▶ Auxiliary trajectories from the same task family

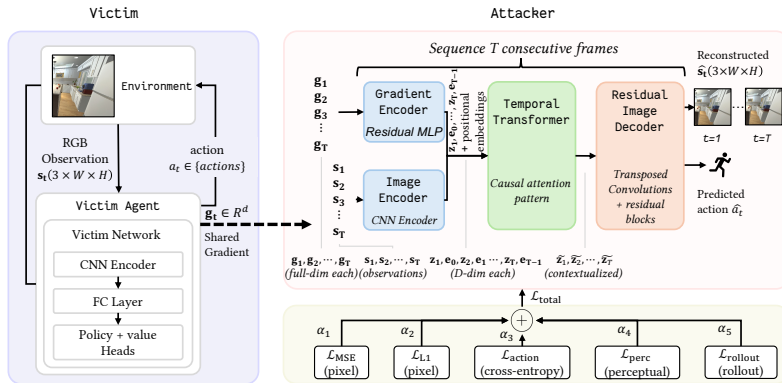
Client (A2C or PPO) sends per-step gradients

$$\mathbf{g}_t = \nabla_{\theta} \left[\ell_t^{\text{pol}}(\theta) + c_v (V_{\theta}(\mathbf{s}_t) - \hat{R}_t)^2 - c_e \mathcal{H}(\pi_{\theta}(\cdot | \mathbf{s}_t)) \right]_{\theta_0}$$

Goal

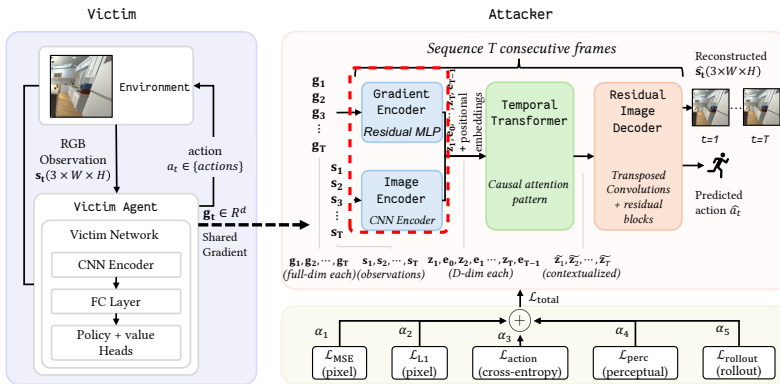
$$(\mathbf{g}_1, \dots, \mathbf{g}_T) \longrightarrow (\hat{\mathbf{s}}_{1:T}, \hat{\mathbf{a}}_{1:T})$$

TRACE: Encode, Contextualize, Decode



- ▶ Encoders turn each gradient and past frame into a 1024-dim token
- ▶ A causal transformer adds trajectory context; a decoder outputs frame and action

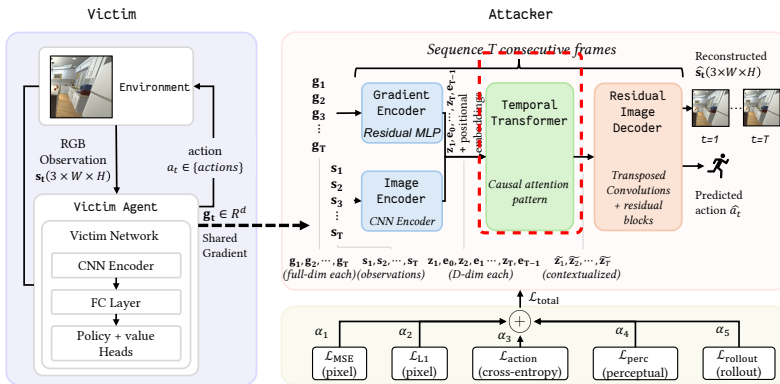
Encoders Compress Each Gradient and Frame



► Residual MLP: $g_t \in \mathbb{R}^{936,102} \rightarrow z_t \in \mathbb{R}^{1024}$

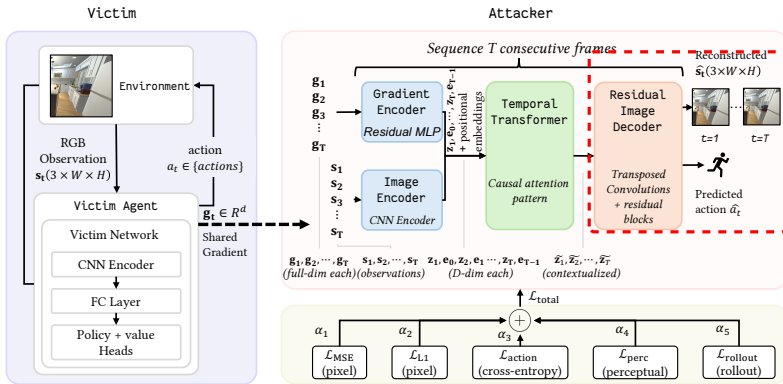
► CNN encodes past frames (true or predicted) into context tokens e_t

A Causal Transformer Carries Scene Context



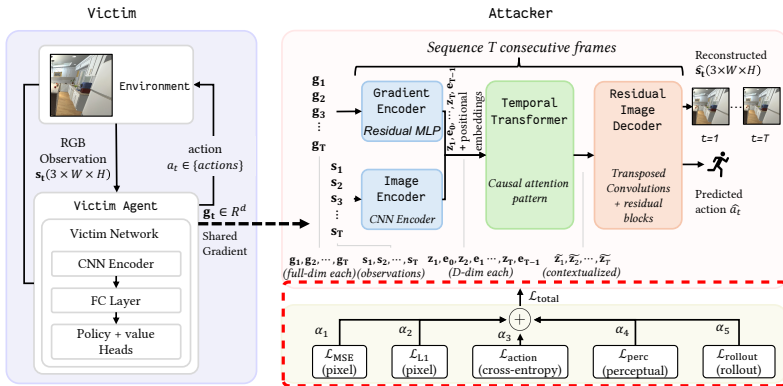
- Causal attention over $[\mathbf{z}_1, \mathbf{e}_0, \mathbf{z}_2, \mathbf{e}_1, \dots, \mathbf{z}_T, \mathbf{e}_{T-1}]$
- Step t sees gradients up to t and earlier frames only

The Decoder Recovers the Frame and the Action



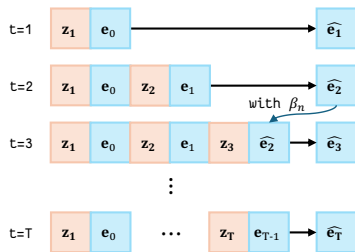
- Image head: transposed convolutions $\rightarrow 84 \times 84$ RGB frame $\hat{s}_t$
- Action head: two-layer MLP $\rightarrow$ softmax over K actions

A Multi-Term Training Objective

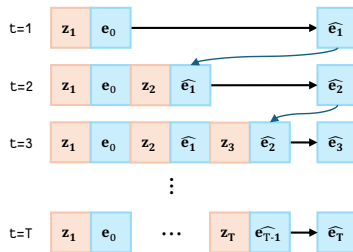


- ▶ $\mathcal{L}_{\text{total}} = \alpha_1 \mathcal{L}_{\text{MSE}} + \alpha_2 \mathcal{L}_{\text{L1}} + \alpha_3 \mathcal{L}_{\text{act}} + \alpha_4 \mathcal{L}_{\text{perc}} + \alpha_5 \mathcal{L}_{\text{rollout}}$
- ▶ Pixel (MSE, L1), perceptual (LPIPS), action (cross-entropy), and rollout terms

Scheduled Sampling Closes the Train–Inference Gap



(a) Training: true frame, or own prediction with prob. β_n



(b) Inference: own predictions only

- ▶ Teacher forcing trains on true frames, but inference sees only its own predictions
- ▶ Scheduled sampling ramps in own predictions; a rollout loss trains on self-generated context

A2C and PPO Send the Same Gradient

Proposition

At θ_0 the PPO ratio is 1, so clipping is inactive and

$$\mathbf{g}_t = \underbrace{-\hat{A}_t \nabla_{\theta} \log \pi_{\theta}(a_t | \mathbf{s}_t)}_{\text{policy}} + \underbrace{2c_v (V_{\theta_0}(\mathbf{s}_t) - \hat{R}_t) \nabla_{\theta} V_{\theta}(\mathbf{s}_t)}_{\text{value}} - \underbrace{c_e \nabla_{\theta} \mathcal{H}(\pi_{\theta}(\cdot | \mathbf{s}_t))}_{\text{entropy}}$$

- ▶ Policy term carries the action; value and entropy terms add state information
- ▶ One attack covers both A2C and PPO

Actions Are Recoverable in Closed Form

Theorem

With ReLU features $\mathbf{h}(\mathbf{s}_t) \geq 0$ and logits $\mathbf{W}_a^\top \mathbf{h}(\mathbf{s}_t)$, the column sums of the policy-head gradient are

$$\sigma_k = \hat{A}_t \|\mathbf{h}(\mathbf{s}_t)\|_1 (\pi_k - \mathbf{1}[k = a_t]) \implies a_t = \arg \max_k |\sigma_k| \quad (K \geq 3)$$

- ▶ No learning needed: read the action straight from the gradient
- ▶ Exact while the entropy weight is small; fails only for a deterministic policy

Temporal Context Never Hurts

Temporal-context gain

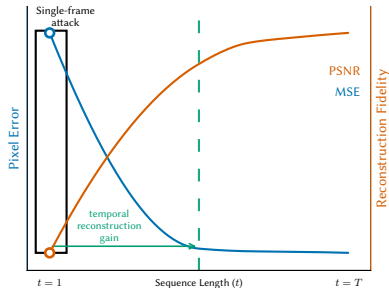
$$\Delta I_t := I(\mathbf{s}_t; \mathbf{g}_{1:t-1} \mid \mathbf{g}_t) \geq 0$$

Theorem

1. $\text{MSE}^*(\mathbf{g}_{1:t}) \leq \text{MSE}^*(\mathbf{g}_t)$
2. Under conditional sufficiency:

$$\Delta I_t = I(\mathbf{s}_t; \mathbf{g}_{1:t-1}) - I(\mathbf{g}_t; \mathbf{g}_{1:t-1})$$

- ▶ Past gradients help with what $\mathbf{g}_t$ does not reveal about $\mathbf{s}_t$
- ▶ GAE breaks (ii), so we test the scaling empirically



Expected trend (illustrative)

Setup: Point-Goal Navigation in AI2-THOR

Victim

- ▶ RGB 84×84 , 5 discrete actions
- ▶ Shared-CNN actor-critic
- ▶ Trained with PPO or A2C

Baselines

- ▶ DLG (Zhu et al., 2019)
- ▶ IG (Geiping et al., 2020)
- ▶ Ltl (Wu et al., 2023)

100 held-out trajectories, $T = 8$; MSE, PSNR, SSIM, LPIPS, action accuracy

TRACE Reconstructs Private Trajectories



Held-out PPO trajectory: ground truth (top), TRACE (bottom)

- ▶ Layout, walls, windows, and counters recovered; fine texture lost
- ▶ All 8 actions recovered

TRACE Beats Optimization and Learned Baselines

Method	PPO					A2C				
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Act. % $\uparrow$	Time	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Act. % $\uparrow$	Time
DLG	5.47	0.043	1.224	16.6	31.0 s	6.11	0.036	1.184	51.0	76.1 s
IG	8.73	0.166	0.766	31.5	39.3 s	9.51	0.093	0.686	32.5	39.6 s
Ltl	16.79	0.529	0.671	100	1.2 ms	17.65	0.508	0.639	100	0.2 ms
TRACE	18.77	0.627	0.362	100	4.5 ms	18.87	0.583	0.358	100	3.0 ms

Means over 100 held-out trajectories, $T = 8$

- ▶ DLG and IG stay below 10 dB and take tens of seconds per frame
- ▶ Over Ltl: +2.0 dB (PPO), +1.2 dB (A2C), LPIPS nearly halved, at ms per frame

The Gain Comes from Temporal Context

T	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
1	12.85	0.513	0.781
8	18.74	0.617	0.375
16	18.57	0.614	0.395
32	19.62	0.634	0.377
64	20.29	0.666	0.375

PPO; ablations use a smaller encoder

- ▶ $T=1 \rightarrow 8$: +5.9 dB PSNR, LPIPS halved
- ▶ Longer context helps further: 20.3 dB at $T=64$
- ▶ At $T=1$, TRACE is below Ltl (16.79 dB): the gain is temporal, not a stronger encoder

Images Leak from Early Layers; Actions Leak Everywhere

Setting	PSNR $\uparrow$	Act. % $\uparrow$
CNN layers only	18.48	100.0
FC layer only	17.72	100.0
Heads only	14.12	99.6
w/o action loss	18.56	19.0
w/o L1 loss	17.90	100.0

One gradient component only, or one loss term removed (PPO, $T = 8$)

- ▶ Heads-only gradients lose 4.6 dB: visual detail sits in early layers
- ▶ Actions stay $\geq 99.6\%$ from any component
- ▶ Without the action loss, accuracy falls to 19% (chance)

Leakage Transfers to Unseen Scenes

$p\%$	PPO		A2C	
	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$
0	18.8	0.362	18.9	0.358
10	20.9	0.273	21.2	0.252
20	22.1	0.224	22.7	0.208
50	23.8	0.161	25.0	0.132

Fine-tuned on $p\%$ of target-scene trajectories; $p = 0$ is zero-shot

- ▶ Zero-shot: 18.8 dB with 100% action accuracy
- ▶ 10% target data adds ~ 2 dB; 50% adds 5–6 dB

Leakage Persists Across Victim Architectures

Victim	K	PSNR $\uparrow$	Act. % $\uparrow$
CNN	10	18.92	99.5
Tiny ViT	5	13.99	99.0
IMPALA CNN	5	21.85	97.4
Wider IMPALA CNN	5	22.36	98.5
CNN+GRU	5	18.62	97.5
Multi-modal CNN	5	19.78	97.4

- ▶ Action accuracy stays $\geq 97\%$ for every victim
- ▶ Images: hardest for Tiny ViT (14.0 dB), easiest for IMPALA CNNs (~ 22 dB)

PPO victims, separate attacker each; K = number of actions

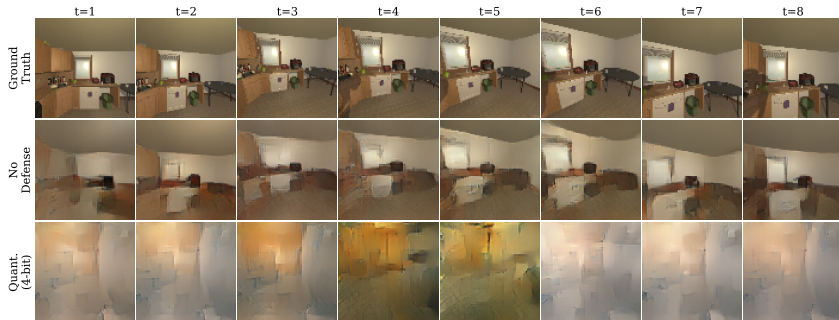
Cheap Defenses Leave Leakage Intact

Defense	PPO		A2C	
	PSNR	Act. %	PSNR	Act. %
None	18.9	99.9	18.4	99.1
Prune, keep 10%	18.8	99.9	18.0	89.0
Quant. 8-bit	18.5	99.9	18.4	98.3
Noise $\sigma=0.01$	18.1	94.9	18.3	96.2
Quant. 4-bit	12.4	37.9	13.3	50.2
Noise $\sigma=0.1$	14.2	59.9	16.5	64.1
DP-SGD	12.1	20.1	13.3	25.2

Zero-shot attacker trained on clean gradients; DP-SGD averaged over three settings

- ▶ Pruning, 8-bit quantization, and small noise barely change the attack
- ▶ 4-bit, $\sigma=0.1$, and DP-SGD suppress it, at an unmeasured utility cost

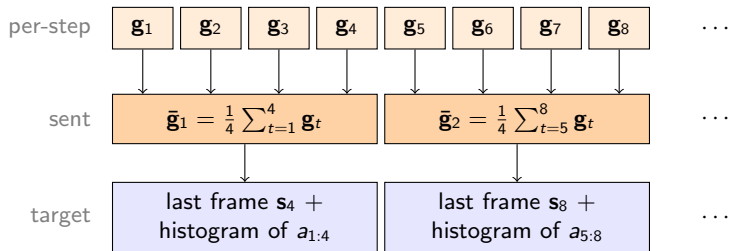
Strong Perturbation Blurs Reconstructions



PPO attacker adapted on 20% of target data

- ▶ No defense: room layout and furniture remain visible
- ▶ 4-bit quantization: scene structure washes out

What If Clients Share Averaged Gradients?

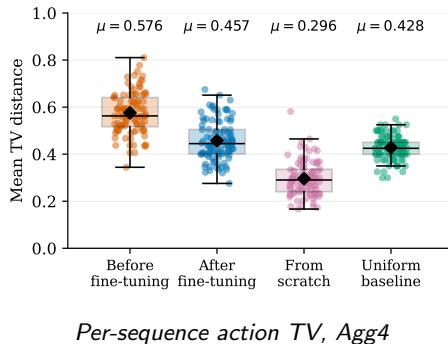


Four-step averaging (Agg4) shown; Agg8 averages eight steps

- Averaging breaks the one-gradient-one-frame link
- Tested with 4- and 8-step windows on the PPO victim

Averaging Weakens but Does Not Remove Leakage

Attacker	PSNR $\uparrow$	LPIPS $\downarrow$	TV $\downarrow$
<i>Agg4</i>			
Before fine-tuning	17.90	0.410	0.576
After fine-tuning	18.04	0.412	0.457
From scratch	17.03	0.510	0.296
Uniform guess	—	—	0.428
<i>Agg8</i>			
From scratch	16.03	0.559	0.268
Uniform guess	—	—	0.304



- ▶ Last frame of each window still leaks at 16–18 dB
- ▶ Only the from-scratch attacker beats a uniform guess on action mix (TV)

For More

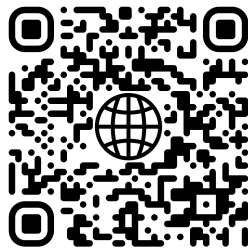
Further details can be found in the paper.



Paper



Code



Web