Anomaly-driven Reinforcement Learning

Saurabh Varshneya[1], Maik Schürmann[2], Philipp Liznerski[1], Mayank Ahuja[1], Jan C. Aurich[2], Sophie Fellenz[1], Marius Kloft[1]

[1]Machine Learning Group, RPTU University Kaiserslautern-Landau
[2]Institute for Manufacturing Technology and Production Systems (FBK), RPTU University Kaiserslautern-Landau

Abstract

Learning additional reward signals—beyond the intrinsic rewards from the environment—can guide the agent more quickly toward optimal policies during training. Recent work suggested training autoregressive models to maximize the likelihood of states observed in expert trajectories. States that are likely then render a positive reward. Such methods closely align with anomaly detection (AD), where the task is to find observations that deviate significantly from typical behavior. We frame reward generation as an AD problem, generalizing existing techniques and proposing a novel methodology for learning reward signals. Specifically, we train state-of-the-art AD models capable of recognizing normal behavior from expert observations while actively learning to detect anomalous behavior from the agent's sub-optimal observations, thus improving the reward learning process. We show that recent autoregressive reward-learning methods are, in fact, implementations of basic AD methods. Empirically, we show the effectiveness of the proposed reward function across various tasks.

Introduction

Reinforcement learning (RL) offers a powerful framework for training agents to solve sequential decision-making problems through trial and error. By interacting with an environment and receiving feedback in the form of rewards, agents can learn to optimize long-term outcomes without requiring explicit supervision.

In many practical scenarios, however, we are not starting from scratch. Expert trajectories are often available in the form of sequential observations such as videos, system states, or sensor logs. Such observations can provide strong guidance to accelerate learning. However, these expert data often lack information about the actions taken or the reward signals that guided the expert's decisions (Bruce et al. 2023; Hoang, Dinh, and Nguyen 2023; Torabi, Warnell, and Stone 2018). This makes it difficult to apply standard RL methods such as imitation learning, which typically assume access to full interaction data (Edwards et al. 2019; Escontrela et al. 2023; Misra et al. 2024).

In such settings, the agent must learn from unlabeled expert observations to recover effective behavior. This presents a unique challenge: how can we leverage expert data without action labels to accelerate reinforcement learning? Prior approaches focus heavily on video data (Bhateja et al. 2023; Pan et al. 2025). Most existing methods learn from unlabeled expert videos, extracting visual dynamics or behavior patterns, but do not generalize beyond visual inputs (Escontrela et al. 2023; Misra et al. 2024). Modern techniques use generative video models as reward surrogates. For example, VIPER (Escontrela et al. 2023) trains a video prediction transformer on expert videos and uses the model's prediction likelihood as a reward to guide policy optimization during online RL. Similarly, VeoRL (Pan et al. 2025) learns latent behavior abstractions from videos to construct a world model that shapes offline policy learning but these approaches remain video-centric. Although powerful, they rely on visual modalities and pretrained video dynamics, limiting their applicability when expert data is comprised of other sequential signals such as sensor readings, keyframes, or logs.

To address this limitation, we propose a novel framework

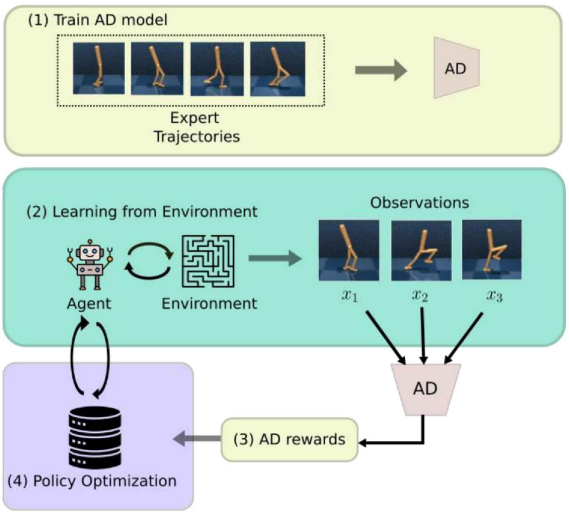

Figure 1: First, we train an AD model to learn a temporal representation of the expert data, which serves as the "normal" behavior. Then, we use the trained AD model to penalize observations from the RL agent that deviate from this learned normal behavior.

based on anomaly detection (AD), which learns reward signals from expert observations by identifying deviations from "normal" behavior. Specifically, we first train an AD model on sequences of expert-only observations to learn a representation of typical dynamics. Then, during reinforcement learning, the agent receives intrinsic rewards that penalize anomalous transitions, encouraging it to generate observation sequences that align closely with the expert data distribution. AD methods for video and non-video data have evolved separately from RL, focusing on identifying undesirable behavior. We argue that when modern AD methods are used in RL, they can provide richer, and better semantically grounded reward signals that capture the structure of both expert and non-expert behavior. By framing reward learning through AD, our approach generalizes recent methods that rely solely on generative video models (such as VIPER or VeoRL) and extends naturally to a wider range of sequential observation types. Unlike these video-centric approaches, our method is modality-agnostic and can operate on structured, non-visual data, including kinesthetic signals, sensor logs, and abstract system states.

Furthermore, we introduce a weakly supervised extension of AD that incorporates both expert and non-expert trajectories during training. This allows the agent to learn more nuanced behavioral boundaries and benefits from partially labeled data, which is common in practical settings but largely underutilized in existing work. To the best of our knowledge, our approach is the first to explore this direction within the context of action-free reinforcement learning.

Through this lens of anomaly-driven RL, our work unifies and extends prior approaches by showing that reward learning from unlabeled expert data can be framed as anomaly detection problem. In the following sections, we demonstrate how this formulation not only matches video-based methods when applied to visual data, but also enables generalization to a broader class of non-visual, sequential RL problems.

Related Work

Learning from Expert Data A substantial amount of research in RL explores the use of expert demonstrations to guide policy learning when explicit reward signals are unavailable. Imitation Learning (IL) is a central framework in this context, seeking to reproduce expert behavior by minimizing the discrepancy between the agent's and the expert's actions. Typical IL approaches, such as Behavioral Cloning (Hussein et al. 2017) and Generative Adversarial Imitation Learning (Ho and Ermon 2016) operate on trajectories consisting of state–action or observation–action pairs to model the expert policy.

In contrast, the proposed method removes this dependency by extracting policy-relevant information directly from observation sequences without requiring access to expert actions.

Generative Modeling for Reward Learning Recent advances in RL have explored generative models as implicit reward estimators from expert demonstrations. Instead of explicitly inferring rewards through imitating expert policies, these approaches learn a generative model over observation sequences and define rewards as the likelihood of the observed future under the learned dynamics model. Video-prediction–based formulations such as Dreamer CURL (Laskin, Srinivas, and Abbeel 2020), and VIPER (Escontrela et al. 2023) exemplify this paradigm, where the agent's reward corresponds to the log-likelihood of predicted future frames given the past. While effective in modeling visual dynamics, such likelihood-based rewards often capture low-level pixel fidelity rather than high-level behavioral semantics, leading to overfitting to visual reconstruction quality instead of behavioral alignment (Huang et al. 2024; Yu et al. 2025).

Anomaly Detection on Time-series and Video data Anomaly detection has been extensively studied in both time-series and video domains, where the objective is to identify deviations from learned patterns of normal behavior in sequence data. In time-series data, reconstruction- and forecasting-based models remain dominant: recurrent and convolutional autoencoders (Malhotra et al. 2016; Xu et al. 2018) and probabilistic variants such as VAEs and VAEs with recurrent decoders (Li et al. 2019; Su et al. 2019) learn to reconstruct or predict normal trajectories, and anomalies are detected through large reconstruction residuals or low reconstruction likelihoods. Recent approaches extend these ideas using attention mechanisms and graph-structured networks to capture cross-channel dependencies and temporal context in multivariate sensor data (Audibert et al. 2020; Zhao, Deng, and Zhou 2022).

In parallel, video anomaly detection has evolved from convolutional autoencoders (Hasan et al. 2016) and ConvLSTM architectures (Lu et al. 2019) to memory-augmented networks such as MemAE (Gong et al. 2019) and MNAD (Park, Noh, and Ham 2020), which reconstruct prototypical normal dynamics while suppressing anomalous patterns. Predictive frameworks that model future frames (Lu et al. 2019; Georgescu et al. 2021) and recent masked-autoencoder or contrastive formulations (Tran, Nguyen, and Kim 2023; Liu et al. 2023) explicitly learn temporal coherence, improving robustness to noise and context shifts on video data.

AD-driven RL framework

We consider an reinforcement learning (RL) setting in which the environment reward is unknown or uninformative. Instead, we are provided with a set of expert observation sequences (videos or outputs from sensors) from which we aim to derive a surrogate reward signal. Let the expert observation sequence be denoted as

$$\{x_1, x_2, \ldots, x_t, \ldots, x_T\},$$

where each $x_t \in \mathcal{O}$ represents the observation frame at time t, and T is the total number of frames in the sequence. The complete expert dataset is given by

$$\mathcal{D}_E = \{\tau_E^{(i)}\}_{i=1}^{N}, \qquad \tau_E^{(i)} = (x_1^{(i)}, \ldots, x_{T_i}^{(i)}).$$

No corresponding action or reward information is available.

Environment model. We assume the underlying environment follows a (partially observed) Markov decision process (POMDP)

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \mathcal{O}, \Omega, \gamma),$$

where $\mathcal{S}$ denotes the latent state space, $\mathcal{A}$ the action space, $P(s_{t+1} \mid s_t, a_t)$ the transition dynamics, $\mathcal{O}$ the observation space, $\Omega(o_t \mid s_t)$ the observation function, and $\gamma \in (0, 1)$ the discount factor. The agent follows a policy $\pi(a_t \mid h_t)$ with history h_t or learned representation of past observations.

Expert normality modeling The AD model is trained directly on expert observations to learn the spatial and temporal regularities that define expert behavior. Each training instance consists of a short temporal window of consecutive frames,

$$x_{t-K:t} = (x_{t-K}, \ldots, x_t),$$

where K denotes the window length. Given such input windows, the AD model A_θ is optimized to accurately model the distribution of normal expert observations through a generic objective of the form

$$\min_\theta \ \mathbb{E}_{I_{t-K:t} \sim \mathcal{D}_E} \big[\mathcal{L}\big(A_\theta(x_{t-K:t}),\, x_{t-K:t}\big) \big],$$

where $\mathcal{L}(\cdot)$ is a task-specific normality loss capturing the reconstruction, prediction, or feature-consistency error between the model's output and the observed sequence. This objective formulation subsumes a broad range of video anomaly detection frameworks, including autoencoding, predictive, and contrastive paradigms. For instance, the suitable training objectives are:

(i) One-class (Deep SVDD) for time-series:

$$\min_\theta \ \mathbb{E}_{\tau \sim \mathcal{D}_E} \sum_t \big\| A_\theta(x_{t-K:t}) - c \big\|_2^2$$

(ii) Density-based (e.g., flow, VAE, diffusion):

$$\min_\theta \ -\mathbb{E}_{\tau \sim \mathcal{D}_E} \ \sum_t \log p_\theta\big(x_{t-K:t}\big)$$

(iii) Predictive consistency (video prediction as normality):

$$\min_\theta \ \mathbb{E}_{\tau \sim \mathcal{D}_E} \ \sum_t \ell\big(x_{t+1}, \hat{x}_{t+1}(x_{t-K:t})\big). \qquad (1)$$

Each formulation induces a distinct anomaly score s_t and, consequently, a different surrogate reward $r_t^{\mathrm{AD}} = g(s_t)$. Density-based formulations (ii) encompass recent video-likelihood approaches such as *VIPER-RL* (Escontrela et al. 2023), where a generative video model estimates $p_\theta(x_t)$ and the reward corresponds to the log-likelihood of the observation under the expert distribution. Predictive objectives (iii) subsume future-frame or reconstruction-based reward shaping, widely adopted in model-based or self-supervised RL (Pathak et al. 2017; Hafner et al. 2020; Wu, Li, and et al. 2023). Thus, the proposed AD-RL framework unifies and generalizes these reward-shaping paradigms under a single principle: learning anomaly-based scores that quantify the alignment between agent observations and expert-consistent dynamics.

In our implementation, we instantiate A_θ with two complementary architectures: *VAD-CLIP* (Wu et al. 2024), which employs contrastive learning in the CLIP feature space to ensure temporal alignment of expert transitions, and *MAE-CvT* (Wu, Li, and et al. 2023), a masked autoencoding framework with a cross-video transformer that reconstructs occluded frame patches to capture spatial–temporal continuity.

Modeling sub-optimal behavior through anomaly detection. While the previous formulation models expert behavior purely from normal observation sequences, we extend this framework to incorporate sub-optimal preferences by framing reward learning as a *semi-supervised anomaly detection* problem. In this setting, expert trajectories are treated as *normal* samples, whereas sub-optimal trajectories are explicitly labeled as *anomalous*. Such sub-optimal data can be readily obtained by executing random or partially trained policies in the environment, or by collecting agent rollouts that terminate prematurely.

Formally, let $\mathcal{D}_E = \{x^{(i)}_{t-K:t}\}$ denote the set of expert windows and $\mathcal{D}_N = \{x^{(j)}_{t-K:t}\}$ the set of non-expert (negative) windows. The semi-supervised objective augments the unsupervised normality loss with an anomaly separation term:

$$\min_\theta \ \mathbb{E}_{x_{t-K:t} \sim \mathcal{D}_E} \big[\mathcal{L}_{\text{norm}}\big(A_\theta(x_{t-K:t}), x_{t-K:t}\big) \big] + \lambda\, \mathbb{E}_{x_{t-K:t} \sim \mathcal{D}_N} \big[\mathcal{L}_{\text{anom}}\big(A_\theta(x_{t-K:t})\big) \big],$$

where $\mathcal{L}_{\text{norm}}(\cdot)$ enforces accurate modeling of expert dynamics, and $\mathcal{L}_{\text{anom}}(\cdot)$ penalizes low anomaly scores for explicitly non-expert samples. The coefficient λ balances the contribution of the two terms. This formulation allows the AD model to jointly refine the notion of "normality" while learning to discriminate sub-optimal behavior, thereby embedding both expert and non-expert preferences directly into the learned reward signal.

Within the semi-supervised anomaly detection framework, the two models, VAD-CLIP and MAE-CvT, capture complementary aspects of sub-optimality. In VAD-CLIP, weak supervision is implemented by aligning video segments with text prompts that describe normal and anomalous events, thus enforcing semantic separation between expert and non-expert behaviors without requiring frame-level annotations. This contrastive alignment in the CLIP embedding space encourages temporal coherence among expert observations while pushing sub-optimal trajectories away from the expert manifold. In contrast, MAE-CVT operates in the reconstruction domain: it learns to accurately reconstruct masked frame patches for expert sequences, while yielding higher reconstruction errors for non-expert samples that violate the spatial–temporal consistency of expert behavior. Together, these two models jointly refine the learned notion of normality—contrastively in the feature space and reconstructively in the pixel space—yielding a more discriminative and stable anomaly-based reward signal.

Anomaly-driven rewards for reinforcement learning. The anomaly detection model offers a natural mechanism for

deriving surrogate rewards in reinforcement learning. Given the per-frame anomaly score $s_t = A_\theta(I_{t-K:t})$, we define the anomaly-driven reward as its inverse:

$$r_t^{\mathrm{AD}} = g(s_t), \qquad g'(s_t) < 0,$$

where a smaller anomaly score (indicating expert-like behavior) yields a higher reward. This reward can be directly integrated into any RL algorithm as a drop-in replacement for the environment reward. To stabilize training, temporal smoothing and robust clipping are optionally applied:

$$\bar{s}_t = (1-\lambda)s_t + \lambda \bar{s}_{t-1}, \qquad \tilde{s}_t = \mathrm{clip}_{[q_\alpha, q_\beta]}(\bar{s}_t),$$

and the corresponding smoothed reward is computed as $r_t^{\mathrm{AD}} = g(\tilde{s}_t)$.

Formally, the objective optimized by the policy π_η is

$$J(\pi_\eta) = \mathbb{E}_{\pi_\eta, P, \Omega}\Big[\sum_{t=0}^{\infty} \gamma^t \, r_t^{\mathrm{AD}}\Big],$$

where maximizing $J(\pi_\eta)$ encourages the agent to align its observation occupancy $\rho_\pi(o)$ with the expert observation distribution $\rho_E(o)$. While AD-based rewards capture expert-likeness effectively, they may not by themselves promote sufficient exploration in high-dimensional environments. In general, this limitation can be alleviated by adding intrinsic exploration bonuses such as curiosity or entropy regularization. Giving the final reward as:

$$r_t = r_t^{\mathrm{AD}} + \beta r_t^{expl}$$

In our implementation, we employ *DreamerV3* as the underlying RL algorithm, which inherently facilitates exploration through its latent world-model dynamics and actor-critic optimization. Thus, no explicit exploration module or intrinsic reward is used; the agent learns solely from the anomaly-driven reward signal.

Experiments

We train two anomaly detection models in a semi-supervised fashion. The first is a Masked Autoencoder with a Convolutional vision Transformer backbone (MAE–CvT) (Ristea et al. 2024; Wu et al. 2021), trained on four DMC vision (Tassa et al. 2018) tasks (walker walk, cheetah run, cartpole balance, and reacher hard) and four Atari (Bellemare et al. 2013) tasks (Boxing, Zaxxon, Atlantis, and Freeway). The second model, VADClip, is trained only on the same four DMC tasks.

In the following, we describe the data preparation, model architecture, training, and inference details for the two AD models. Subsequently, we report their performance on AD and RL tasks.

Using MAE-CvT as a Reward Function

Data Preparation We collect observations from expert demonstrations (expert trajectories) and a random agent (random trajectories), treating frames from expert trajectories as normal data and those from random trajectories as anomalous data. We then construct training data using normal frames and synthetic frames, where we extract objects from anomalous frames and embed them onto the normal frames. The synthetic frames account for five percent of the training dataset.

Model Architecture We modify the original MAE-CvT (Ristea et al. 2024) architecture by removing the self-distillation branch and the classification head. Furthermore, we remove the product of the reconstruction loss and gradients in the objective, and replace it with their sum.

Semi-supervised MAE–CvT We train on a mix of clean frames and synthetic-abnormal frames. For synthetic-abnormal samples, the input is the composite image (made from the frame of a random observation pasted on an expert observation), while the target is the corresponding clean frame with an extra mask that defines the composition. For clean samples, the input equals the target (the composition mask is zero). MAE is trained to reconstruct clean targets from both clean and synthetic-abnormal inputs, thereby learning to differentiate between abnormal and normal frames. To learn the temporal context, image differences between contiguous frames (referred to as gradients) are used in conjunction with the reconstruction loss.

Training objective At time t, the we have input $x_t \in \mathbb{R}^{H\times W\times C}$, and define the motion image $G_t = |x_{t+3} - x_{t-3}|$. We divide an image into patches, where each patch is denoted by p_i (index t can be ignored without the loss of generality). Let P_i be the set of co-ordinates of pixels for patch p_i, and $\ell_i = \|\hat{p}_i - p_i\|_2^2$, where $\hat{p}_i$ is the reconstructed patch. We mask the patches for MAE with a mask $m_i \; in\{0,1\}$. We compute the gradient for each patch and normalize it as follows:

$$g_i = \frac{1}{|P_i|} \sum_{(u,v)\in P_i} G_t(u,v),$$

$$w_i = \frac{g_i + \varepsilon}{\sum_{j=1}^{L}(g_j + \varepsilon)},$$

$\mathcal{L}_{\mathrm{raw}} = \frac{\sum_i m_i \, \ell_i}{\sum_i m_i + \varepsilon}$ is the masked auto-encoder reconstruction loss.

$$\mathcal{L} \;=\; \mathcal{L}_{\mathrm{raw}} \;+\; \lambda \sum_{i=1}^{L} w_i \, \ell_i, \qquad s_t \;=\; \mathcal{L}\,.$$

Anomaly score (inference) We mirror training: compute the difference between the input and predicted patch ℓ_i, take the masked average for the $\mathcal{L}_{raw}$ term and the gradient-normalized weighted sum for the second term, thus compute the score as:

$$s_t \;=\; \mathcal{L}_{raw} \;+\; \lambda \sum_i w_i \, \ell_i$$

Hyperparameters AdamW with cosine LR unless noted. *DMC:* 64×64, patch p=8, in-ch= 9, out-ch= 4, encoder depth 24, decoder depth 8, heads 16, mask ratio ρ=0.50, batch 16, 18 epochs, λ=1.0, α=1, ε=10^{-3}, gradient pooling $p\times p$, temporal gap Δ=3.
Atari: frames resized to 96×96 RGB, p=8, same in/out channels and CvT/decoder, λ=1.0, α=1, ε=10^{-3}, Δ=3.

Training VAD-CLIP

We train a CLIP-style video–text model on weak labels at the clip level. DMC frames are fed to the visual encoder and short textual prompts to the text encoder.

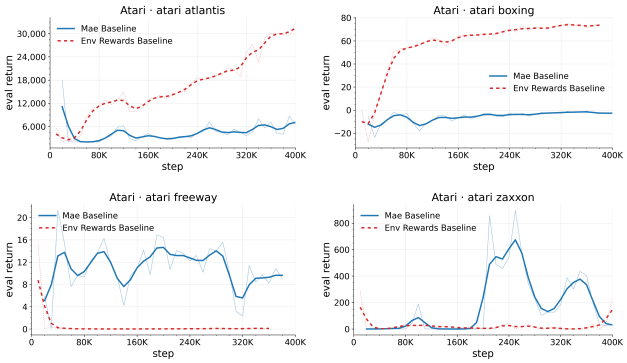

Figure 2: Atari: evaluation return vs. steps. MAE–CvT (blue) vs. environment-reward oracle (red).

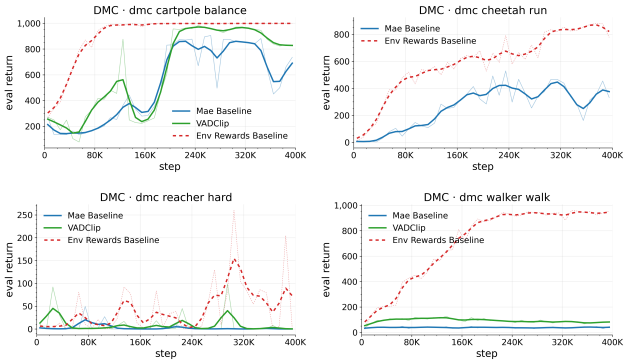

Figure 3: DMC: evaluation return vs. steps. MAE–CvT (blue) vs. oracle (red); VADClip where present (green).

Inputs/prompts. Frames are resized to the CLIP visual resolution and encoded frame-wise; text prompts are minimal class phrases built from the dataset's label map (we do not apply any learnable/prefix prompts).

Backbone The model follows CLIP, which consists of a visual encoder (based on ResNet or ViT) and a Transformer text encoder. Both encoders map to a shared embedding space where cosine similarity acts as a classifier.

Training objective Let $s_{t,c}$ be the cosine-similarity logit at frame t for class c, and let $\mathcal{T}_k \subset \{1, \dots, L\}$ be the indices of the top-k frames (per class) in the clip. We aggregate per-class evidence by

$$\bar{s}_c = \frac{1}{k} \sum_{t \in \mathcal{T}_k} s_{t,c}.$$

Two weakly-supervised losses are combined:

$$\begin{aligned} \mathcal{L}_{\text{bin}} &= \text{BCE}\left(\sigma\left(\frac{1}{k} \sum_{t \in \mathcal{T}_k} s_{t,\text{anom}} \right),\ y \in \{0,1\} \right), \\ \mathcal{L}_{\text{mil}} &= -\sum_{c} \tilde{y}_c \log \frac{\exp(\bar{s}_c)}{\sum_{c'} \exp(\bar{s}_{c'})}. \end{aligned} \tag{2}$$

where y is the clip's binary normal/anomalous label and $\tilde{y}$ is a normalized multi-label target vector (when available). A small regularizer encourages separation between "normal" and the anomalous text embeddings. The total loss is

$$\mathcal{L} = \mathcal{L}_{\text{bin}} + \mathcal{L}_{\text{mil}} + \eta\, \mathcal{R}_{\text{text}}, \quad \eta \ll 1\,.$$

Anomaly score (inference) Given a test clip, we compute per-frame logits $s_{t,c}$ against the text prompts. A clip-level anomaly score is the top-k mean of the anomaly logit $\bar{s}_{\text{anom}}$ (binary path), and a multi-class score is obtained from $\{\bar{s}_c\}_{c \neq \text{normal}}$ (MIL path). For per-frame visualization we use

$$a_t = \max_{c \neq \text{normal}} s_{t,c}.$$

These scores are used for AUC/AP evaluation with the same top-k temporal pooling as in training.

Semi-supervised VADClip In VAD-CLIP, weak supervision arises from aligning video segments with text prompts that describe normal and anomalous events, thus enforcing the difference between expert and non-expert behaviors.

Evaluation of AD performance

We evaluated the performance of the AD methods on Atari and DMC vision datasets, where normal data comprises the test split of normal trajectories, and anomalous data comprises the test split of the random trajectories. Table 1 1 shows that MAE-CvT attains near-perfect Micro and Macro AUC on DMC vision datasets, whereas on Atari, MAE-CvT shows sub-optimal results (see Table 2). One of the reasons for poor AD performance could be the diverse observation space in Atari. This could be improved by increasing model capacity and hyperparameter tuning. Table 3 shows the performance of VAD-CLIP on the DMC vision tasks. The model exhibits near-perfect AUCs here as well, indicating that it effectively differentiates between normal and anomalous frames.

Table 1: DMC — frame-level AUC of MAE–CvT anomaly score.

Task	Clips	Macro AUC	Micro AUC
walker_walk	3	0.998	0.998
cheetah_run	41	0.930	0.958
cartpole_balance	33	0.999	0.999
reacher_hard	24	1.000	0.999

Table 2: Atari — frame-level AUC of MAE–CvT anomaly score.

Task	Clips	Macro AUC	Micro AUC
boxing	26	0.495	0.478
zaxxon	3	0.421	0.454
atlantis	6	0.633	0.607
freeway	26	0.477	0.517

Table 3: DMC — frame-level AUC of VAD Clip Anomaly Scores.

Task	AUC	AP
walker_walk	0.9295	0.2792
cheetah_run	0.9654	0.6127
cartpole_balance	0.9110	0.2499
reacher_hard	0.8298	0.3004

Evaluation of AD-driven RL agents

To compare the different AD-based reward functions, we trained the RL agent using AD-based rewards and ignoring the environment rewards. For evaluation, we computed the environment rewards that are accumulated by the agent. Figure 2 shows that the modified MAE-CvT model achieves superior returns as compared to environment returns in 2 of the 4 tasks. MAE-CvT and VADClip achieve comparable results as compared to environment returns in the DMC vision tasks (see Figure 3).

Conclusion

In this work, we demonstrated how recent advancements in anomaly detection (AD) can be effectively integrated into reinforcement learning (RL) through reward shaping. We further showed that several existing reward-shaping paradigms can be unified under our general AD-driven RL framework, providing a principled perspective on leveraging normality and deviation signals as reward functions. Our experimental analysis reveals that AD-based rewards are particularly advantageous in environments with sparse or delayed feedback, such as Atari tasks.

However, we also observed that using random trajectories as anomalous samples is suboptimal: although the AD model can easily discriminate expert from random behaviors, it provides weak reward gradients for intermediate or partially successful trajectories, leading to reduced environment returns. These findings suggest that constructing richer anomalous datasets—by including trajectories that represent system failures or terminal states—can improve the learned AD representations and produce more informative reward signals. Future work will explore adaptive strategies for curating such anomalous experiences and extending the AD-driven reward formulation to more complex, multi-agent, and continuous-control settings.

Acknowledgments

Part of this work was conducted within the DFG Research Unit FOR 5359 on Deep Learning on Sparse Chemical Process Data (BU 4042/2-1, KL 2698/6-1, and KL 2698/7-1). MK and SF further acknowledge support by the DFG TRR 375 (ID 511263698), the DFG SPP 2298 (KL 2698/5-2), and the DFG SPP 2331 (FE 2282/1-2, FE 2282/6-1, and KL 2698/11-1). Additional support was provided by the Carl-Zeiss Stiftung within the initiatives AI-Care and Process Engineering 4.0, as well as the BMBF award 01—S2407A.

References

Audibert, J.; Michiardi, P.; Guyard, F.; Marti, S.; and Zuluaga, M. A. 2020. USAD: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3395–3404.

Bellemare, M. G.; Naddaf, Y.; Veness, J.; and Bowling, M. 2013. The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research*, 47: 253–279.

Bhateja, C.; Guo, D.; Ghosh, D.; Singh, A.; Tomar, M.; Vuong, Q.; Chebotar, Y.; Levine, S.; and Kumar, A. 2023. Robotic offline rl from internet videos via value-function pre-training. *arXiv preprint arXiv:2309.13041*.

Bruce, J.; Anand, A.; Mazoure, B.; and Fergus, R. 2023. Learning About Progress From Experts. In *The Eleventh International Conference on Learning Representations*.

Edwards, A.; Sahni, H.; Schroecker, Y.; and Isbell, C. 2019. Imitating Latent Policies from Observation. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 1755–1763. PMLR.

Escontrela, A.; Adeniji, A.; Yan, W.; Jain, A.; Peng, X. B.; Goldberg, K.; Lee, Y.; Hafner, D.; and Abbeel, P. 2023. Video prediction models as rewards for reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 68760–68783.

Georgescu, M. I.; Ionescu, R. T.; Khan, F. S.; and Shah, M. 2021. Anomaly detection in video via self-supervised and multi-task learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12742–12752.

Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M. R.; Venkatesh, S.; and Hengel, A. v. d. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1705–1714.

Hafner, D.; Lillicrap, T.; Norouzi, M.; and Ba, J. 2020. Dream to Control: Learning Behaviors by Latent Imagination. In *ICLR*.

Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A. K.; and Davis, L. S. 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 733–742.

Ho, J.; and Ermon, S. 2016. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29.

Hoang, M.-H.; Dinh, L.; and Nguyen, H. 2023. Learning from Pixels with Expert Observations. arXiv:2306.13872.

Huang, T.; Jiang, G.; Ze, Y.; and Xu, H. 2024. Diffusion reward: Learning rewards via conditional video diffusion. In *European Conference on Computer Vision*, 478–495. Springer.

Hussein, A.; Gaber, M. M.; Elyan, E.; and Jayne, C. 2017. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2): 1–35.

Laskin, M.; Srinivas, A.; and Abbeel, P. 2020. Curl: Contrastive unsupervised representations for reinforcement learning. In *International conference on machine learning*, 5639–5650. PMLR.

Li, D.; Chen, D.; Jin, B.; Shi, L.; Goh, J.; and Ng, S.-K. 2019. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In *Proceedings of the 28th International Conference on Artificial Neural Networks (ICANN)*, 703–716.

Liu, Y.; Liu, X.; Zhang, X.; Li, Y.; Song, S.; Li, H.; and Liu, G. H. 2023. Self-distilled masked autoencoders are efficient video anomaly detectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lu, Y.; Li, J.; Shi, C.; Jia, F.; Yan, W.; Chen, X.; and Li, X. 2019. Future frame prediction for anomaly detection–A new baseline. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6536–6545.

Malhotra, P.; Ramakrishnan, A.; Anand, G.; Vig, L.; Agarwal, P.; and Shroff, G. 2016. LSTM-based encoder-decoder for multi-sensor anomaly detection. In *Proceedings of the 33rd International Conference on Machine Learning Workshop on Anomaly Detection*.

Misra, D.; Saran, A.; Xie, T.; Lamb, A.; and Langford, J. 2024. Towards principled representation learning from videos for reinforcement learning. *arXiv preprint arXiv:2403.13765*.

Pan, M.; Zheng, Y.; Li, J.; Wang, Y.; and Yang, X. 2025. Video-Enhanced Offline Reinforcement Learning: A Model-Based Approach. *arXiv preprint arXiv:2505.06482*.

Park, H.; Noh, J.; and Ham, B. 2020. Learning memory-guided normality for anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14372–14381.

Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-driven Exploration by Self-supervised Prediction. In *ICML*.

Ristea, N.-C.; Georgescu, M.-I.; Ionescu, R. T.; Khan, F. S.; and Shah, M. 2024. Self-Distilled Masked Auto-Encoders are Efficient Video Anomaly Detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Su, Y.; Zhao, Y.; Niu, C.; Liu, R.; Sun, W.; and Pei, D. 2019. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2828–2837.

Tassa, Y.; Doron, Y.; Muldal, A.; Erez, T.; Li, Y.; Casarini, F.; Kumar, A.; Mann, T.; Szita, I.; Erez, T.; and Lillicrap, T. 2018. DeepMind Control Suite. *arXiv preprint arXiv:1801.00690*.

Torabi, F.; Warnell, G.; and Stone, P. 2018. Generative Adversarial Imitation from Observation. *CoRR*, abs/1807.06158.

Tran, S.; Nguyen, H.; and Kim, J. 2023. MAE-CVT: Masked autoencoders for continuous video transformers in anomaly detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3006–3015.

Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; and Zhang, L. 2021. CvT: Introducing Convolutions to Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Wu, P.; Zhou, X.; Pang, G.; Zhou, L.; Yan, Q.; Wang, P.; and Zhang, Y. 2024. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6074–6082.

Wu, Y.; Li, K.; and et al. 2023. Self-Distilled Masked Autoencoders are Efficient Video Anomaly Detectors. In *CVPR*.

Xu, H.; Chen, W.; Zhao, N.; Li, Z.; Bu, J.; Li, Z.; Liu, Y.; Zhao, Y.; and Pei, D. 2018. Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. In *Proceedings of the 2018 World Wide Web Conference*, 187–196.

Yu, R.; Wan, S.; Wang, Y.; Gao, C.-X.; Gan, L.; Zhang, Z.; and Zhan, D.-C. 2025. Reward Models in Deep Reinforcement Learning: A Survey. *arXiv preprint arXiv:2506.15421*.

Zhao, R.; Deng, W.; and Zhou, Z. 2022. Multivariate time-series anomaly detection via graph attention networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(5): 1986–1998.