ATTENTION ON DEMAND: SURPRISE-DRIVEN SELECTIVE ATTENTION FOR REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention mechanisms are widely used in reinforcement learning (RL) to capture long-range temporal dependencies, yet many architectures compute temporal attention at every timestep regardless of local predictability. In partially observable environments, recurrent representations may suffice during predictable transitions, whereas longer-range temporal context can become more useful when local dynamics fail to explain incoming observations. We introduce Attention-on-Demand (AoD), a surprise-conditioned framework that adaptively weights recurrent and attention representations. AoD combines a recurrent core with causal multi-head attention and uses an auxiliary forward model to predict the next latent observation. The resulting latent prediction error serves as a local surprise signal that conditions a differentiable gate between the two representations. Across the evaluated partially observable benchmarks, AoD achieves episodic returns comparable to full-attention baselines. In the primary MemoryS7 evaluation, its mean gate activation is $\mathbb{E}[\omega_t] = 0.646$, corresponding to 35.4% lower representational attention reliance relative to a full-attention reference. Gate activation increases from 0.38 to 0.85 across surprise quintiles and also rises under observation corruption. These results show that a surprise-conditioned gate can vary the attention representation's contribution with local prediction error.

1 INTRODUCTION

Self-attention effectively models long-range dependencies, but its computational cost grows quadratically with sequence length (Keles et al., 2023; Pan et al., 2022). Transformer-based RL architectures such as GTrXL (Parisotto et al., 2020) and Decision Transformer (Chen et al., 2021) employ attention as a fixed architectural component, although long-range temporal context may not be equally informative at every timestep. While local dynamics may suffice for portions of a trajectory (Sutton et al., 1998), partial observability can make additional historical context important (Kaelbling et al., 1998). This motivates a central question: *when should a policy rely more strongly on attention than on recurrent memory?*

Existing approaches address related but distinct problems. Efficient attention methods such as Reformer Kitaev et al. (2020) and Performer Choromanski et al. (2020) reduce attention complexity but do not regulate its relative contribution according to temporal predictability. Similarly, mixture-of-experts methods Shazeer et al. (2017) route information among specialized subnetworks but do not explicitly regulate recurrent versus attention representations using prediction error.

A central hypothesis of this work is that prediction error can provide a temporally grounded signal for regulating attention reliance. Predictive models capture discrepancies between expected and observed dynamics (Rao & Ballard, 1999; Ha & Schmidhuber, 2018b; Hafner et al., 2019a), while partial observability makes temporal memory important (Kaelbling et al., 1998; Igl et al., 2018). Recurrent models provide compact temporal representations (Hausknecht & Stone, 2015; Wierstra et al., 2007), whereas attention provides access to longer-range context. These complementary properties motivate using prediction error to regulate their relative contribution.

We introduce **Attention-on-Demand (AoD)**, illustrated in Figure 1, a surprise-conditioned adaptive-gating framework that treats attention reliance as a state-dependent representational quantity. AoD combines a recurrent branch with a latent prediction head and a causal multi-head attention branch for longer-range temporal reasoning. Unlike standard adaptive fusion that derives mixing weights

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

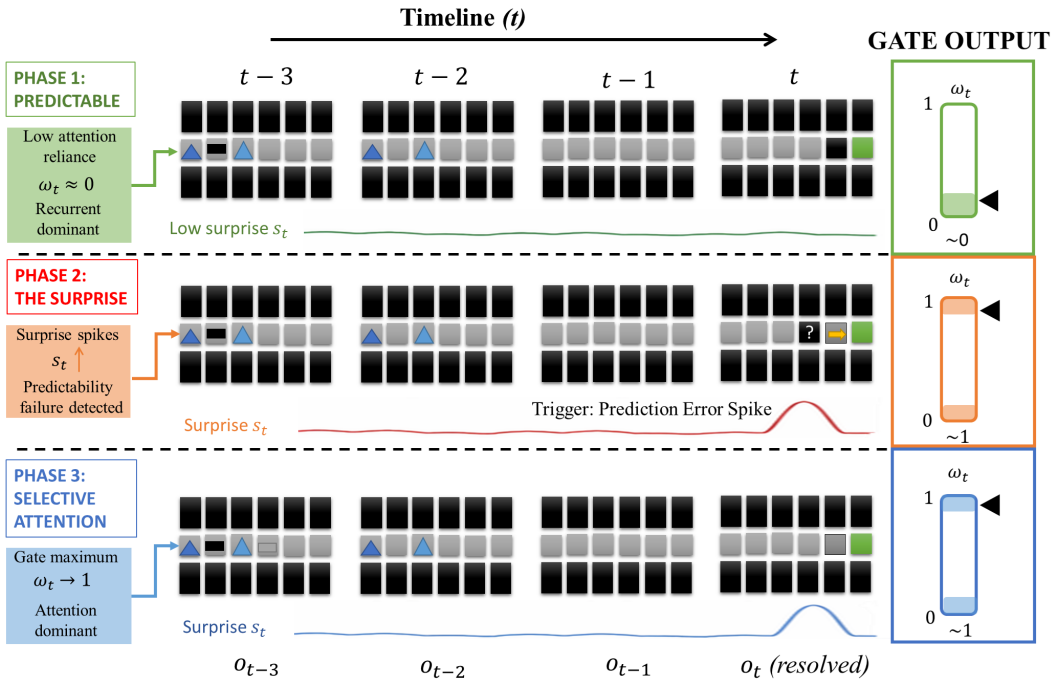

Figure 1: **Attention-on-Demand (AoD): surprise-conditioned adaptive representation.** *AoD dynamically modulates the relative contribution of recurrent and attention representations using latent prediction error.* **Phase 1 (Predictable):** Low surprise s_t favors the recurrent representation. **Phase 2 (Surprise):** Increased prediction error signals reduced local predictability. **Phase 3 (Increased Attention Reliance):** A larger gate value ω_t increases the contribution of longer-range attention. Both branches execute at every timestep; the gate controls their relative representational contribution rather than conditional execution.

from representation features alone, AoD uses latent prediction error as an explicit measure of local prediction failure to condition the gate. The resulting surprise-conditioned representational routing links local predictability to temporal-context reliance, enabling the policy to adaptively balance recurrent memory and longer-range attention according to the temporal reasoning demands of each state.

Empirically, we compare AoD with recurrent, fixed-mixture, adaptive-gating without surprise, and full-attention baselines. We evaluate three questions: whether AoD maintains task performance relative to full attention, whether gate activation varies systematically with prediction surprise, and whether explicit surprise input contributes beyond adaptive fusion alone. We further examine gate behavior under observation corruption to test whether attention reliance changes as observations become less predictable.

Our main contributions are: (1) a surprise-conditioned gating mechanism that uses latent prediction error to regulate the relative contribution of recurrent and attention representations; (2) a dual-branch RL architecture that combines compact recurrent memory with causal temporal attention; and (3) empirical evidence that attention reliance varies with prediction surprise and observation uncertainty while AoD maintains performance comparable to full-attention baselines.

2 Related Work

2.1 Attention and Recurrent Memory in Reinforcement Learning

Attention mechanisms Vaswani et al. (2017) provide a means of modeling long-range temporal dependencies in sequential decision-making. Transformer-based approaches include GTrXL for online

Table 1: **Illustrative results motivating AoD's recurrent and attention branches.** Scores use each source's reported metric and are comparable only within the same benchmark. Higher is better.

Benchmark	Method	Architecture	Score	Metric	Reference
POPGym	MLP	Feedforward	0.067	Normalized MMER (↑)	Morad et al. (2023)
	Lin. Attn (FART)	Attention	0.138	Normalized MMER (↑)	Morad et al. (2023)
	LSTM	Recurrent	0.255	Normalized MMER (↑)	Morad et al. (2023)
	GRU	**Recurrent**	**0.349**	**Normalized MMER (↑)**	Morad et al. (2023)
DMLab-30	Deep LSTM	Recurrent	99.3	Mean HNR (↑)	Parisotto et al. (2020)
	GTrXL	**Attention**	**117.6**	**Mean HNR (↑)**	Parisotto et al. (2020)

MMER = max-mean episodic reward; HNR = human-normalized return. Scores are comparable only within each benchmark.

RL Parisotto et al. (2020), Decision Transformer for offline RL Chen et al. (2021), Gato for generalist policy modeling from offline trajectories Reed et al. (2022), and TrMRL for meta-RL Melo (2022). In these approaches, attention is part of the model architecture rather than a representation whose relative contribution is explicitly conditioned on latent prediction error at each timestep. Efficient-attention methods such as Longformer Beltagy et al. (2020), BigBird Zaheer et al. (2020), and Reformer Kitaev et al. (2020) reduce the cost of processing long sequences through sparse or structured attention, but do not use prediction error to weight recurrent and attention representations.

Recurrent architectures such as GRUs Cho et al. (2014) and LSTMs Hochreiter & Schmidhuber (1997) provide compact temporal memory and remain effective for partially observable sequential decision-making Smirnov & Gu (2025); Pedamonti et al. (2025). Alternative sequence architectures, including temporal convolutions Bai et al. (2018) and structured state-space models Gu & Dao (2023); Smith et al. (2023); Lu et al. (2023), provide other mechanisms for capturing temporal dependencies. POPGym further demonstrates the importance of explicitly evaluating memory mechanisms under partial observability Morad et al. (2023).

Table 1 summarizes representative temporal architectures relevant to AoD. Reported performance is retained in the metric used by each original study; values from different benchmarks should therefore not be compared directly.

Architectural Motivation. Table 1 presents results from separate studies that motivate examining both recurrent and attention-based temporal representations. On POPGym, GRU obtains an aggregate normalized MMER of 0.349, exceeding the LSTM (0.255) and FART (0.138) entries shown in the table Morad et al. (2023). In a separate DMLab-30 evaluation, GTrXL (GRU) achieves a mean human-normalized return of 117.6, compared with 99.3 for the LSTM baseline Parisotto et al. (2020). Because these results come from different benchmarks and experimental settings, they motivate AoD's combination of a GRU and a causal attention branch rather than establishing which representation is generally superior. AoD learns their relative contribution as a function of the current observation, recurrent state, and latent prediction error.

2.2 Hierarchical and Mixture-of-Experts RL

Hierarchical reinforcement learning Dietterich (2000) decomposes decision-making into temporally or behaviorally distinct components Sutton et al. (1999); Bacon et al. (2017), while mixture-of-experts (MoE) models Jacobs et al. (1991) learn to select or combine specialized experts. Subsequent MoE approaches investigate expert specialization and parameter scaling Eigen et al. (2013); Lepikhin et al. (2020); Fedus et al. (2022); Obando-Ceron et al. (2024). The cited hierarchical methods address behavioral decomposition, while the MoE methods route among learned experts. AoD instead weights recurrent and attention representations within a single policy, using latent prediction error alongside the observation encoding and recurrent state to condition the gate.

2.3 Adaptive Computation and Dynamic Routing

Adaptive computation methods vary the amount or location of computation according to the input Graves (2016), including spatially adaptive processing, dynamic depth, and early exits Figurnov et al. (2017); Bolukbasi et al. (2017); Dehghani et al. (2018); Elbayad et al. (2019). In partially observable RL, the current observation may not contain sufficient information about the underly-

ing state, making temporal memory useful Hausknecht & Stone (2015). AoD is related to adaptive computation through its learned gate, but its current soft-gating formulation evaluates both branches at every timestep and adapts their *representational contribution* rather than conditionally skipping computation. The gate incorporates latent prediction error alongside the current observation encoding and recurrent state.

2.4 Predictive Models and World Models in RL

Model-based RL approaches such as Dreamer (Hafner et al., 2019a) and PlaNet (Hafner et al., 2019b) learn predictive latent dynamics for representation learning and decision-making (Hafner et al., 2023; Ha & Schmidhuber, 2018a). Prediction-based intrinsic signals have also been used for exploration, including forward-model prediction error (Pathak et al., 2017), random-network prediction error (Burda et al., 2018), and ensemble disagreement (Sekar et al., 2020). AoD uses prediction error differently: rather than treating it as an intrinsic reward, AoD uses it to condition the relative weighting of recurrent and attention representations.

Summary of Distinction. Prior work has explored attention-based policies Parisotto et al. (2020); Chen et al. (2021), recurrent memory Hausknecht & Stone (2015); Wierstra et al. (2007), efficient attention Kitaev et al. (2020), expert routing Fedus et al. (2022), and prediction-based exploration Pathak et al. (2017). AoD combines recurrent and attention representations within an RL policy and explicitly uses latent prediction error, together with the observation encoding and recurrent state, to condition their relative weighting. This distinguishes its surprise-conditioned representational gate from the mechanisms discussed above.

3 Methodology

Motivated by the complementary roles of recurrent memory and temporal attention discussed in the preceding section, we develop AoD to use latent prediction error when weighting their representations within a policy.

3.1 Problem Formulation

We consider a partially observable Markov decision process (POMDP) Kaelbling et al. (1998), specified by $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \Omega, \mathcal{O}, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $\mathcal{P}$ is the transition model, $\mathcal{R}$ is the reward function, Ω is the observation space, $\mathcal{O}$ is the observation model, and $\gamma \in [0, 1)$ is the discount factor. At timestep t, the agent observes $o_t \in \Omega$ and selects $a_t \in \mathcal{A}$. The interaction history $h_t = (o_0, a_0, o_1, a_1, \ldots, a_{t-1}, o_t)$ induces a belief over the latent state $\xi_t \in \mathcal{S}$, $b_t(\xi) = p(\xi_t = \xi \mid h_t)$ Kaelbling et al. (1998).

AoD does not explicitly maintain this belief. Instead, its learned representations summarize the observation sequence $o_{0:t} = (o_0, \ldots, o_t)$ for action selection. The objective is to learn a policy $\pi_\theta(a_t \mid o_{0:t})$ that maximizes expected discounted return:

$$\max_\theta \ \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} \gamma^t r_t\right], \tag{1}$$

where r_t is the reward at timestep t and T is the final timestep of an episode. The policy implements this dependence through a learned representation z_t of $o_{0:t}$, defined below.

3.2 Overview of Surprise-Conditioned Adaptive Gating

We introduce **Attention-on-Demand (AoD)**, a surprise-conditioned adaptive-gating framework that treats attention reliance as a state-dependent quantity. As illustrated in Figure 2, AoD maintains a recurrent representation that summarizes observation history and an attention representation with direct access to a recent context window. A learned gate weights these representations using the current encoding, recurrent state, and a latent prediction-error signal Rao & Ballard (1999).

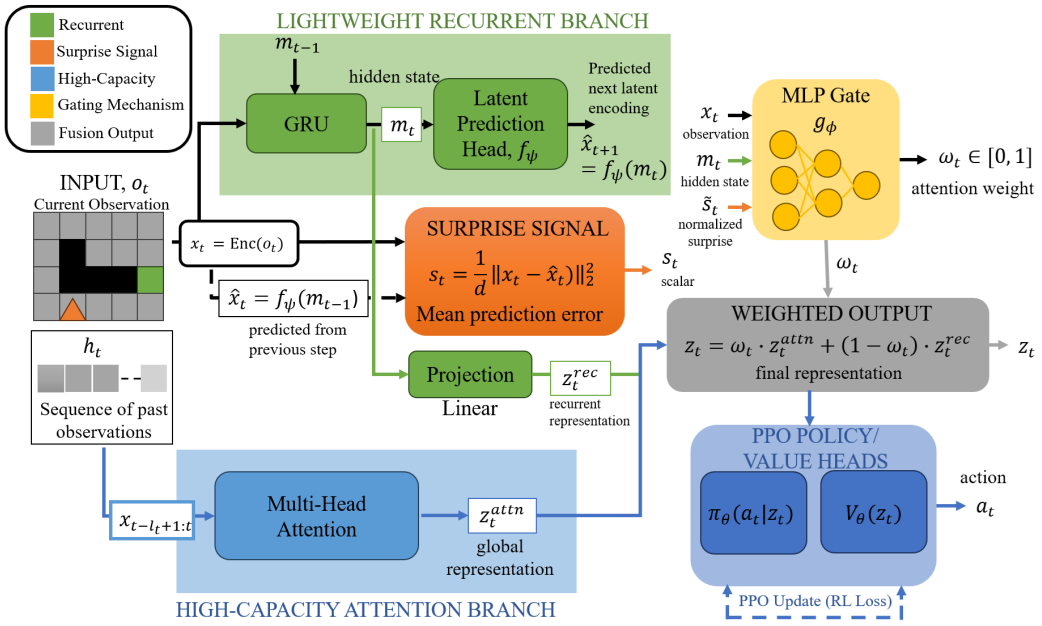

Figure 2: **Attention-on-Demand (AoD) architecture.** A One-Hot Encoder maps observation o_t to latent encoding x_t, which is processed by recurrent and causal-attention branches. The discrepancy between the current encoding x_t and the prediction formed from the preceding recurrent state m_{t-1} defines surprise s_t. The normalized surprise $\tilde{s}_t$, current encoding x_t, and updated recurrent state m_t condition the gate, which produces $\omega_t \in [0, 1]$. The gate combines z_t^{rec} and z_t^{attn} into z_t, which feeds the policy and value heads. Both branches execute at every timestep.

3.3 Dual-Branch Representation

Each observation is converted to a one-hot representation and mapped to the shared latent space:

$$x_t = \text{Enc}(o_t), \qquad x_t \in \mathbb{R}^d,$$

where Enc denotes the one-hot observation encoding followed by a learned linear/MLP projection to latent dimension d. Subsequent computation operates on the encoded observation x_t.

The recurrent branch computes

$$m_t = \text{GRU}(x_t, m_{t-1}), \tag{2}$$
$$z_t^{\text{rec}} = W^{\text{rec}} m_t, \tag{3}$$

where $m_t \in \mathbb{R}^d$ is the recurrent hidden state and $W^{\text{rec}} \in \mathbb{R}^{d \times d}$ is a learned projection matrix Cho et al. (2014).

The attention branch operates over the most recent $\ell_t = \min(L, t+1)$ encoded observations, with a maximum context length of $L = 16$:

$$Q_t = \text{CausalAttn}(x_{t-\ell_t+1}, \ldots, x_t), \tag{4}$$
$$q_t = Q_t^{(\ell_t)}, \tag{5}$$
$$z_t^{\text{attn}} = W^{\text{attn}} q_t, \tag{6}$$

where Q_t contains the position-wise outputs of causal multi-head self-attention, q_t is its output at the current position, and $W^{\text{attn}} \in \mathbb{R}^{d \times d}$ is a learned projection matrix. We use learned positional embeddings. Both branch representations lie in the shared latent space, $z_t^{\text{rec}}, z_t^{\text{attn}} \in \mathbb{R}^d$. The attention context contains observations from the current episode only and is reset at episode boundaries.

3.4 Surprise Signal from Prediction Error

To estimate local predictability, the recurrent state predicts the next encoded observation:

$$\hat{x}_{t+1} = f_\psi(m_t), \tag{7}$$

where $m_t \in \mathbb{R}^d$ is the recurrent hidden state and f_ψ is a two-layer prediction MLP with parameters ψ. Thus, the prediction available at timestep t is $\hat{x}_t = f_\psi(m_{t-1})$. Surprise is the mean squared latent prediction error:

$$s_t = \frac{1}{d} \left\| x_t - \mathrm{sg}(\hat{x}_t) \right\|_2^2, \tag{8}$$

where $x_t = \mathrm{Enc}(o_t) \in \mathbb{R}^d$ and $\mathrm{sg}(\cdot)$ stops gradients through the stored prediction. The prediction head is trained through $\mathcal{L}_{\mathrm{pred}}$ (Section 3.6).

For scale stability, surprise is normalized as

$$\tilde{s}_t = \frac{s_t - \mu_s}{\sigma_s + \epsilon}, \tag{9}$$

where μ_s and σ_s are running estimates of the surprise mean and standard deviation, and $\epsilon > 0$ is a numerical-stability constant.

3.5 Surprise-Conditioned Gating

AoD conditions attention reliance on the encoded observation, recurrent state, and normalized surprise:

$$\omega_t = g_\phi(x_t, m_t, \tilde{s}_t), \qquad \omega_t \in [0, 1], \tag{10}$$

where g_ϕ is a two-layer gating MLP with parameters ϕ and a sigmoid output. The final representation is

$$z_t = \omega_t z_t^{\mathrm{attn}} + (1 - \omega_t) z_t^{\mathrm{rec}}, \tag{11}$$

where z_t^{attn} and z_t^{rec} are the attention and recurrent representations, respectively, and ω_t is their learned mixing weight. Small ω_t favors the recurrent branch, whereas large ω_t increases attention contribution. The fused representation z_t is provided to the policy $\pi_\theta(a_t \mid z_t)$ and value function $V_\theta(z_t)$.

3.6 Training Objective

We minimize the PPO objective Schulman et al. (2017):

$$\mathcal{L}_{\mathrm{PPO}} = \mathcal{L}_{\mathrm{clip}} + c_1 \mathcal{L}_V - c_2 \mathcal{H}(\pi_\theta), \tag{12}$$

where $\mathcal{L}_{\mathrm{clip}}$ is the negative clipped policy surrogate, $\mathcal{L}_V$ is the value loss, and $\mathcal{H}(\pi_\theta)$ is the policy entropy. The coefficients c_1 and c_2 weight the value and entropy terms, respectively. Advantages are estimated using GAE Schulman et al. (2015). AoD additionally minimizes next-latent prediction error and attention reliance:

$$\mathcal{L}_{\mathrm{pred}} = \mathbb{E}_t[\|x_{t+1} - \hat{x}_{t+1}\|_2^2], \qquad \mathcal{L}_{\mathrm{attn}} = \mathbb{E}_t[\omega_t], \tag{13}$$

where $\mathbb{E}_t$ denotes expectation over timesteps, $x_{t+1} = \mathrm{Enc}(o_{t+1})$ is the encoded next observation, $\hat{x}_{t+1} = f_\psi(m_t)$ is its prediction from recurrent state m_t, and $\omega_t \in [0, 1]$ is the attention mixing weight. Episode-boundary transitions are excluded from $\mathcal{L}_{\mathrm{pred}}$.

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{PPO}} + \alpha \mathcal{L}_{\mathrm{pred}} + \lambda \mathcal{L}_{\mathrm{attn}}, \tag{14}$$

where α and λ control the prediction and attention-reliance regularization strengths, respectively.

3.7 Training and Computational Cost

Both branches are evaluated at every timestep. For hidden dimension d and context length $L = 16$, the recurrent update costs $\mathcal{O}(d^2)$ per timestep. Recomputing self-attention over the context window costs $\mathcal{O}(Ld^2 + L^2 d)$ per timestep, including the token projections and attention operation. Therefore, $\mathbb{E}[\omega_t]$ describes representational attention reliance; it does not measure reductions in computation, latency, or energy.

4 EXPERIMENTS

We evaluate Attention-on-Demand (AoD) on partially observable reinforcement learning tasks along three objectives: **(a) Task Performance**, comparing AoD with recurrent, fixed-mixture, and full-attention baselines; **(b) Attention Reliance**, measuring its average representational reliance on attention alongside task performance; and **(c) Adaptive Gating**, examining how gate activation varies with prediction surprise and observation masking.

4.1 ENVIRONMENTS

We evaluate AoD on three MiniGrid environments: `MemoryS7`, `DoorKey`, and `DynamicObstacles`, covering partial observability, memory, and sequential decision-making Chevalier-Boisvert et al. (2023). For the observation-corruption analysis on `MemoryS7`, we replace each observation with zeros at a timestep with probability $p \in \{0, 0.1, 0.3, 0.5\}$; $p = 0$ is the uncorrupted reference.

4.2 BASELINES

We compare AoD with an **MLP**, a **GRU**, a **Transformer (Full Attention)**, and a **Fixed Mixture** ($\omega_t = 0.5$). The AoD w/o Surprise ablation described in Section 4.7 removes surprise from the gate input. All models use $d = 128$ and the same PPO settings.

4.3 EVALUATION METRICS

Table 4 summarizes the evaluation metrics. Episodic return measures task performance. For AoD, $\mathbb{E}[\omega_t]$ is the average attention mixing weight, and $100(1 - \mathbb{E}[\omega_t])\%$ expresses its reduction relative to a unit-weight full-attention reference. We characterize the association between surprise s_t and gate activation ω_t using Spearman's ρ and Pearson's r.

4.4 IMPLEMENTATION DETAILS

All models are trained with Proximal Policy Optimization (PPO) for 2×10^6 environment steps using eight parallel environments, 128-step rollouts, four optimization epochs per update, a mini-batch size of 256, and a learning rate of 3×10^{-4}. Recurrent and attention-based models use a hidden dimension of $d = 128$, with four attention heads and a sequence length of 16 where applicable. We evaluate MLP, GRU, Transformer, Fixed Mixture, and AoD on MemoryS7, DoorKey, and DynamicObstacles across five independent random seeds. For AoD, we additionally evaluate $\lambda \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$ and observation-corruption probabilities $p \in \{0, 0.1, 0.3, 0.5\}$. Unless otherwise stated, results are reported as mean $\pm$ standard deviation across five seeds. All experiments were conducted on a workstation with an Intel Core i7-14700K CPU, NVIDIA GeForce RTX 4080 GPU, and 64 GB DDR4 RAM. Evaluation uses the final checkpoint obtained after 2×10^6 environment steps. Observation corruption is applied only at evaluation time by replacing the current observation with zeros independently with probability $p \in \{0, 0.1, 0.3, 0.5\}$.

Tables 2 and 3 summarize the experimental protocol, including the common PPO configuration, benchmark models and environments, and evaluation metrics. All models are trained under consistent settings across five random seeds, while AoD is additionally evaluated through attention reliance, prediction surprise, and their relationship to characterize its adaptive gating behavior.

4.5 PERFORMANCE AND ATTENTION RELIANCE

We compare AoD's episodic returns with the full-attention baseline and measure its representational reliance on attention using mean gate activation $\mathbb{E}[\omega_t]$. Relative to a unit-weight full-attention reference, the reliance reduction is $R_{\text{attn}} = 100(1 - \mathbb{E}[\omega_t])\%$.

As shown in Table 5, AoD and the full-attention Transformer both obtain a mean return of 0.49 on MemoryS7 and 0.80 on DoorKey. On DynamicObstacles, their means are 0.75 and 0.74, respectively. On MemoryS7, AoD's evaluation mean gate activation of 0.646 corresponds to 35.4% lower

Table 2: Experimental configuration used across all models.

Parameter	Value
Training steps	2×10^6
Parallel environments	8
Rollout length	128
PPO epochs/update	4
Minibatch size	256
Learning rate	3×10^{-4}
Discount factor γ	0.99
GAE parameter	0.95
PPO clipping range	0.2
Hidden dimension d	128
Attention heads	4
Sequence length	16
Prediction-loss weight α	0.1
Default attention penalty λ	0.01
Attention penalty λ	$\{.001, .005, .01, .05, .1\}$
Corruption probability p	$\{0, .1, .3, .5\}$

Table 3: Experimental design for the primary MiniGrid evaluation.

Study	Configuration	Runs
Benchmark	5 M $\times$ 3 E $\times$ 5 S	75
No-surprise ablation	3 E $\times$ 5 S	15
Reg.	4 add. $\lambda \times$ 5 S	20
Corr.	2 M $\times$ 3 add. $p \times$ 5 S	30
Total		**140**

Benchmark includes the default $\lambda = 0.01$ and $p = 0$ conditions. M = models; E = environments; S = seeds; Reg. = regularization; Corr. = corruption; add. = additional.

Table 4: Metrics used to assess AoD.

Metric	Purpose
Return	Task performance
$\mathbb{E}[\omega_t]$	Attention reliance
$1 - \mathbb{E}[\omega_t]$	Reliance reduction
Surprise s_t	Local unpredictability
$\rho(s_t, \omega_t)$	Surprise–gate relation

Table 5: **Main performance and attention-reliance results.** Mean episodic return is reported across five independent seeds for all architectures. For AoD, mean gate activation $\mathbb{E}[\omega_t]$ and relative attention-reliance reduction $R_{\text{attn}} = 100(1 - \mathbb{E}[\omega_t])$ are additionally reported. Episodic returns are reported as mean $\pm$ standard deviation.

Model	MemoryS7	DoorKey	DynamicObs.	$\mathbb{E}[\omega_t]$	$\mathbf{R}_{\text{attn}}$
MLP	0.32 ± 0.08	0.45 ± 0.06	0.38 ± 0.07	–	–
GRU	0.49 ± 0.05	0.76 ± 0.05	0.70 ± 0.06	–	–
Transformer	0.49 ± 0.04	0.80 ± 0.04	0.74 ± 0.05	1.000	0.0%
Fixed Mixture	0.49 ± 0.06	0.78 ± 0.05	0.72 ± 0.06	0.500	50.0%
AoD (Ours)	0.49 ± 0.05	0.80 ± 0.06	0.75 ± 0.06	0.646	35.4%

For Transformer, $\mathbb{E}[\omega_t] = 1$ denotes the full-attention reference; Fixed Mixture uses $w = 0.5$. AoD's reported gate activation is the recorded evaluation mean. R_{attn} measures representational reliance. Gate statistics for AoD are from the default MemoryS7 evaluation.

representational attention reliance relative to the unit-weight reference. Both branches still execute at every timestep.

4.6 Performance–Attention Reliance Trade-off

On MemoryS7, we evaluate AoD at five attention-penalty strengths, $\lambda \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$, across five seeds; $\lambda = 0.01$ is the default. We compare episodic return and mean gate activation $\mathbb{E}[\omega_t]$, with reliance reduction defined as $R_{\text{attn}} = 100(1 - \mathbb{E}[\omega_t])\%$ relative to full attention. Table 7 reports the results.

4.7 Ablation Studies

Attention Regularization. We vary λ while holding other settings fixed to characterize the attention reliance trade-off.

Fixed Mixing and Surprise Conditioning. We compare AoD with a Fixed Mixture ($\omega_t = 0.5$) and an adaptive *AoD w/o Surprise* variant whose gate receives x_t and m_t but not $\tilde{s}_t$. The latter retains the prediction head, auxiliary loss, and other training settings, isolating the effect of supplying surprise to the gate.

As shown in Table 6, Full AoD has higher mean return than AoD w/o Surprise on all three tasks, with differences of 0.02, 0.03, and 0.05 on MemoryS7, DoorKey, and DynamicObstacles, respectively.

Table 6: **Fixed vs. adaptive mixing and ablation of prediction surprise.** Mean episodic return ($\pm$ sample standard deviation) across five independent seeds ($n = 5$).

Model	MemoryS7	DoorKey	DynamicObs.
Fixed Mixture ($\omega = 0.5$)	0.49 ± 0.06	0.78 ± 0.05	0.72 ± 0.06
AoD w/o Surprise	0.47 ± 0.06	0.77 ± 0.06	0.70 ± 0.06
AoD (Full)	$\mathbf{0.49 \pm 0.05}$	$\mathbf{0.80 \pm 0.06}$	$\mathbf{0.75 \pm 0.06}$

The Fixed Mixture provides a non-adaptive reference. These are descriptive five-seed comparisons; the table does not establish that the differences are statistically significant.

4.8 Analysis of Gating Behavior

Surprise and Observation Uncertainty. We examine the relationship between prediction surprise s_t and gate activation ω_t using Spearman correlation (with Pearson as a secondary measure) and surprise quintiles. We further evaluate AoD and the full-attention Transformer under observation corruption $p \in \{0, 0.1, 0.3, 0.5\}$, measuring performance for both models and $\mathbb{E}[\omega_t]$ to determine whether attention reliance increases as observations become less predictable or reliable.

Temporal Dynamics and Gate Stability. We analyze ω_t over the complete 2×10^6-step training trajectory and across evaluation states to assess temporal adaptation and gate stability. Variation in ω_t indicates state-dependent modulation, whereas concentration near a single value indicates behavior approaching a fixed mixture. Detailed gating analyses are provided in the supplementary material.

4.9 Broader Impact

AoD demonstrates that attention reliance can be reduced while maintaining performance comparable to full attention, motivating future conditional execution. Although the current soft-gating formulation evaluates both branches at every timestep, future implementations could run recurrent processing continuously and invoke attention only when triggered, using stored temporal representations as context. This could translate reduced attention reliance into lower inference cost, latency, and energy consumption, particularly for resource-constrained edge and robotic systems.

Limitations and Future Work. While our evaluation spans MiniGrid, long-horizon Passive T-Maze ($L \leq 500$), and heterogeneous POPGym tasks, several questions remain. First, AoD uses soft representational routing, so both pathways are evaluated at every timestep. Consequently, reductions in $\mathbb{E}[\omega_t]$ reflect reduced attention reliance rather than realized FLOP, latency, or energy savings. Converting this behavior into computational savings requires hard or conditional routing, which we leave for future work. Second, although AoD is evaluated with attention contexts up to $L_{\text{ctx}} = 505$, its behavior in higher-dimensional continuous-control settings and rich visual environments remains unexplored. Extending surprise-conditioned routing to such domains is an important direction for future work.

5 Conclusion

We introduced **Attention-on-Demand (AoD)**, a surprise-conditioned adaptive-gating framework that combines recurrent and attention representations in partially observable reinforcement learning. AoD uses latent prediction error as a surprise signal to regulate their state-dependent relative contribution. Across the evaluated benchmarks, AoD maintains performance comparable to full attention. On MemoryS7, its mean gate activation of $\mathbb{E}[\omega_t] = 0.646$ corresponds to 35.4% lower representational attention reliance relative to the full-attention reference. Our analyses further characterize how gating behavior varies with prediction surprise, regularization strength, and observation corruption. Together, these results show that AoD can adjust the relative contribution of recurrent and attention representations while maintaining task performance.

AI USE STATEMENT

Generative AI tools assisted with LaTeX typesetting, mathematical notation checks, phrasing and copyediting, table formatting, and the implementation and debugging of analysis and plotting scripts. They also assisted in organizing and providing feedback on proposed ablations and sensitivity analyses. All AI-assisted text and code were reviewed by the authors, and AI-assisted software was inspected, tested, and validated against empirical logs. Generative AI was not used to generate synthetic datasets, formulate the research hypotheses or theoretical framework, derive mathematical claims or proofs, execute experiments, or independently interpret experimental results. The authors verified all reported results and conclusions and take full responsibility for the final content, claims, mathematical rigor, code, and artifacts presented in this work.

ETHICS STATEMENT

This work uses simulated reinforcement learning benchmarks and does not involve human subjects or personally identifiable information. The authors are responsible for accurate reporting and reproducibility.

REPRODUCIBILITY STATEMENT

The Methodology section describes the architecture and objective, and the Experiments section reports the training and evaluation settings. The appendix provides additional ablations, gating analyses, and per-seed results. An anonymized repository containing the code, configurations, checkpoints, and evaluation scripts is available at `https://anonymous.4open.science/r/AOD-174E/`.

REFERENCES

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Tolga Bolukbasi, Joseph Wang, Ofer Dekel, and Venkatesh Saligrama. Adaptive neural networks for efficient inference. In *International Conference on Machine Learning*, pp. 527–536. PMLR, 2017.

Yuri Burda, Harrison Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and J K Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *Advances in neural information processing systems*, 36:73383–73394, 2023.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1724–1734, 2014.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

Thomas G Dietterich. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of artificial intelligence research*, 13:227–303, 2000.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. Depth-adaptive transformer. *arXiv preprint arXiv:1910.10073*, 2019.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Mikhail Figurnov, Maxwell D Collins, Yukun Zhu, Li Zhang, Jonathan Huang, Dmitry Vetrov, and Ruslan Salakhutdinov. Spatially adaptive computation time for residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1039–1048, 2017.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in Neural Information Processing Systems*, 31, 2018a.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018b.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pp. 2555–2565. PMLR, 2019b.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Matthew J Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable mdps. In *AAAI fall symposia*, volume 45, pp. 141, 2015.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for pomdps. In *International conference on machine learning*, pp. 2117–2126. PMLR, 2018.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. On the computational complexity of self-attention. In *International conference on algorithmic learning theory*, pp. 597–619. PMLR, 2023.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

Dmitry Lepikhin, Hyoukseok Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Chris Lu, Yannick Schroecker, Albert Gu, Emilio Parisotto, Jakob Foerster, Satinder Singh, and Feryal Behbahani. Structured state space models for in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36:47016–47031, 2023.

Luiz C Melo. Transformers are meta-reinforcement learners. In *International Conference on Machine Learning*, pp. 15340–15359. PMLR, 2022.

Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, and Amanda Prorok. Popgym: Benchmarking partially observable reinforcement learning. *arXiv preprint arXiv:2303.01859*, 2023.

Tianwei Ni, Michel Ma, Benjamin Eysenbach, and Pierre-Luc Bacon. When do transformers shine in rl? decoupling memory from credit assignment. *Advances in Neural Information Processing Systems*, 36:50429–50452, 2023.

Johan Obando-Ceron, Ghada Sokar, Timon Willi, Clare Lyle, Jesse Farebrother, Jakob Foerster, Gintare Karolina Dziugaite, Doina Precup, and Pablo Samuel Castro. Mixtures of experts unlock parameter scaling for deep rl. *arXiv preprint arXiv:2402.08609*, 2024.

Xuran Pan, Chunjiang Ge, Rui Lu, Shiji Song, Guanfu Chen, Zeyi Huang, and Gao Huang. On the integration of self-attention and convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 815–825, 2022.

Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Seb Noury, et al. Stabilizing transformers for reinforcement learning. In *International conference on machine learning*, pp. 7487–7498. PMLR, 2020.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, pp. 2778–2787. PMLR, 2017.

Dabal Pedamonti, Samia Mohinta, Martin V Dimitrov, Hugo Malagon-Vina, Stephane Ciocchi, and Rui Ponte Costa. Hippocampus supports multi-task reinforcement learning under partial observability. *Nature Communications*, 16(1):9619, 2025.

Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International Conference on Machine Learning*, pp. 8583–8592. PMLR, 2020.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Ivan Smirnov and Shangding Gu. Rlbenchnet: The right network for the right reinforcement learning task. *arXiv preprint arXiv:2505.15040*, 2025.

Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *International Conference on Learning Representations (ICLR)*, 2023.

Richard S Sutton, Andrew G Barto, and Andrew Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Daan Wierstra, Alexander Foerster, Jan Peters, and Juergen Schmidhuber. Solving deep memory pomdps with recurrent policy gradients. In *International conference on artificial neural networks*, pp. 697–706. Springer, 2007.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

A Appendix

A.1 Supplementary Experimental Analysis

This appendix provides supplementary analyses supporting the main empirical results. We report training dynamics, statistical comparisons, per-seed results, attention-reliance and surprise-conditioned gate analyses, and observation-corruption experiments. We further evaluate AoD under extended memory horizons using Passive T-Maze and across heterogeneous partially observable tasks from POPGym. Together, these analyses provide additional evidence for the performance, robustness, and adaptive representational routing behavior of AoD.

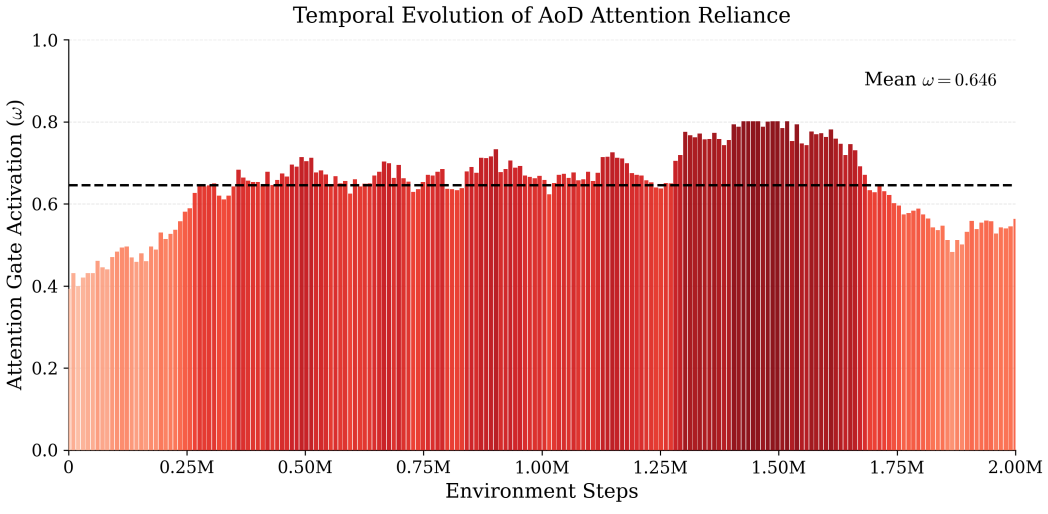

Figure 3: **Temporal evolution of attention reliance in AoD.** The gate activation ω controls the relative contribution of the attention pathway to the fused representation, with lower values indicating greater reliance on the lightweight recurrent pathway and higher values indicating increased use of global temporal context. Bar intensity provides an additional visual encoding of gate magnitude, ranging from light red for lower attention reliance to dark red for higher reliance. The dashed horizontal line denotes the average gate activation over the training trajectory. Each bar represents a per-checkpoint mean gate activation.

Temporal Attention Reliance. Figure 3 shows how mean gate activation changes across training checkpoints. Lower checkpoint means indicate less average contribution from the attention representation, while higher means indicate more. This figure characterizes the evolution of the learned gate during training; state-level variation is examined separately in the surprise-conditioned gate analysis.

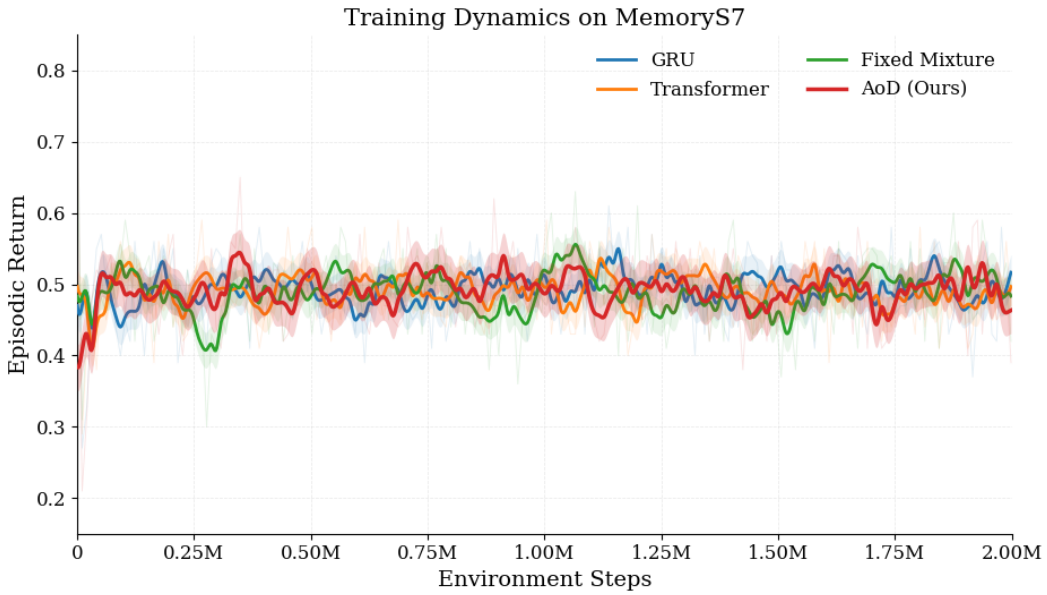

Figure 4: **Training dynamics on MemoryS7 over 2M environment steps.** Episodic return is shown for the GRU, Transformer, Fixed Mixture, and Attention-on-Demand (AoD) architectures for a representative training run (seed 0). Curves are obtained using a five-checkpoint centered moving average followed by cubic-spline interpolation for visualization, while the faint traces show the underlying logged returns. The methods show overlapping ranges of logged episodic return, with AoD exceeding the other curves during some portions of this representative run. Their ordering changes during training; this single-seed visualization does not establish consistent superiority for any architecture.

Full Training Dynamics. Figure 4 provides a representative view of learning behavior over the complete 2M-step training horizon. AoD follows a similar overall return trajectory to the GRU, Transformer, and Fixed Mixture baselines, with periods in which its episodic return rises above those of the other architectures. However, the relative ordering changes repeatedly during training, suggesting that no single architecture consistently dominates this representative run. The figure provides a qualitative view of this representative run; aggregate performance comparisons use the five-seed evaluation.

Surprise–Gate Dynamics. Figure 5 examines whether the learned gating behavior is associated with the prediction-surprise signal that motivates AoD. Across the logged training checkpoints, surprise and gate activation exhibit a positive association, with Pearson $r = 0.70$ and Spearman $\rho = 0.68$. Thus, checkpoints with larger prediction error tend to exhibit greater attention contribution. Because these statistics summarize training checkpoints, the state-level analysis in Table 9 and Figure 6 provides complementary evidence of surprise-conditioned gating during evaluation.

Performance–Attention Reliance Trade-off. Table 7 characterizes the effect of attention regularization on task performance and attention reliance. Increasing λ is associated with progressively lower mean gate activation, indicating reduced contribution from the attention pathway. At lower and moderate regularization strengths, this reduction is accompanied by relatively small changes in episodic return, whereas stronger regularization produces larger reductions in attention reliance

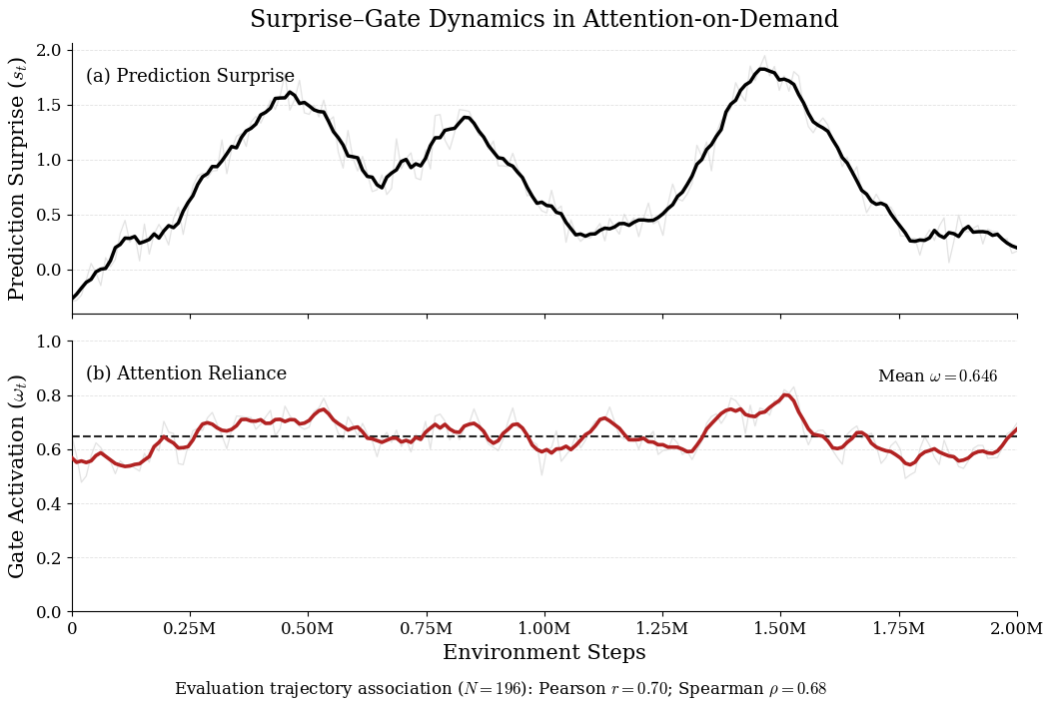

Figure 5: **Surprise–gate dynamics of Attention-on-Demand (AoD) over 2M environment steps on `MemoryS7`.** (a) Prediction surprise s_t, derived from the latent prediction error of the recurrent pathway, and (b) attention gate activation ω_t for a representative training run (seed 0). Faint traces denote the logged values, while solid curves show five-checkpoint centered moving averages for visualization. The dashed line in (b) indicates the mean gate activation, $\bar{\omega} = 0.646$. Across the $N = 196$ logged training checkpoints, checkpoint-mean surprise and gate activation are positively associated (Pearson $r = 0.70$; Spearman $\rho = 0.68$). This training-level association complements the evaluation-state analysis in Table 9 and Figure 6 (Pearson $r = 0.74$; Spearman $\rho = 0.79$).

Table 7: **Performance–attention reliance trade-off under varying attention regularization strengths.** The full-attention Transformer serves as the reference with $\mathbb{E}[\omega_t] = 1.000$ and 0.0% reduction in attention reliance. For AoD, reliance reduction is computed relative to full attention as $100(1 - \mathbb{E}[\omega_t])$. Values report mean $\pm$ standard deviation across five independent random seeds ($n = 5$).

Model / Regularization (λ)	**Return (↑)**	$\mathbb{E}[\omega_t]$	**Reliance Reduction (↑)**
Full-Attention Transformer	0.49 ± 0.05	1.000	0.0%
AoD ($\lambda = 0.001$)	0.50 ± 0.04	0.820 ± 0.022	18.0%
AoD ($\lambda = 0.005$)	0.50 ± 0.05	0.720 ± 0.028	28.0%
AoD ($\lambda = 0.010$, default)	0.49 ± 0.05	0.646 ± 0.029	35.4%
AoD ($\lambda = 0.050$)	0.46 ± 0.06	0.530 ± 0.031	47.0%
AoD ($\lambda = 0.100$)	0.40 ± 0.07	0.420 ± 0.029	58.0%

Values are mean $\pm$ standard deviation across five independent seeds ($n = 5$). Gate statistics are computed from evaluation trajectories for each seed and subsequently aggregated across seeds. Reliance reduction is computed from the corresponding mean gate activation.

together with a greater decrease in task performance. These results characterize a trade-off between task performance and attention reliance rather than a single universally optimal operating point.

Robustness Under Observation Corruption. Table 8 examines model behavior under increasing observation masking using our default regularizer strength ($\lambda = 0.010$). Episodic return degrades gracefully for both AoD and the full-attention Transformer as corruption probability p rises. Con-

Table 8: Robustness to observation masking on MemoryS7. Episodic return and AoD attention reliance across varying observation masking probabilities p evaluated at the default regularizer strength ($\lambda = 0.010$). At each timestep, observations are replaced with zero vectors with probability p ($p = 0.0$ denotes uncorrupted evaluation, identically matching Tables 10 and 7). Values report mean $\pm$ sample standard deviation across five independent random seeds ($n = 5$).

Masking (p)	**AoD Return** (↑)	**Transformer Return** (↑)	**AoD** $\mathbb{E}[\omega_t]$
0.0	0.49 ± 0.05	0.49 ± 0.05	0.646 ± 0.029
0.1	0.46 ± 0.06	0.45 ± 0.06	0.684 ± 0.026
0.3	0.41 ± 0.07	0.39 ± 0.08	0.758 ± 0.030
0.5	0.35 ± 0.08	0.32 ± 0.09	0.822 ± 0.032

AoD exhibits a widening descriptive margin over the full-attention baseline as corruption severity increases ($+0.00 \to +0.01 \to +0.02 \to +0.03$), accompanied by an adaptive shift toward the attention pathway ($\mathbb{E}[\omega_t] : 0.646 \to 0.822$).

Table 9: **Attention reliance conditioned on prediction surprise.** Evaluation timesteps ($N = 196$ per seed) are partitioned into quintiles by latent prediction error. Across-seed statistics report mean gate activation $\mathbb{E}[\omega_t]$, across-seed sample standard deviation ($n = 5$), 95% confidence intervals ($t_{0.975,4} = 2.776$), and low-gate activation rate ($\omega_t < \tau = 0.05$).

Surprise	**Range**	**Gate** $\mathbb{E}[\omega_t]$	**95% CI**	**Low-G Rate** ($\omega_t < 0.05$)
Minimal	0–20%	0.38 ± 0.04	$[0.33, 0.43]$	$4.1\% \pm 1.2\%$
Low	20–40%	0.54 ± 0.03	$[0.50, 0.58]$	$0.8\% \pm 0.5\%$
Moderate	40–60%	0.67 ± 0.03	$[0.63, 0.71]$	$0.0\% \pm 0.0\%$
High	60–80%	0.79 ± 0.03	$[0.75, 0.83]$	$0.0\% \pm 0.0\%$
Extreme	80–100%	0.85 ± 0.02	$[0.83, 0.87]$	$0.0\% \pm 0.0\%$
Range ($\Delta\omega = +0.47$)		*Overall Mean* $\mathbb{E}[\omega_t] = 0.646$*; Pearson* $r = 0.74$*; Spearman* $\rho = 0.79$		

Values report mean $\pm$ sample standard deviation across five independent random training seeds ($n = 5, df = 4$). Within-tier timestep standard deviations ($0.14, 0.12, 0.12, 0.11, 0.09$) reflect intra-trajectory dispersion, whereas reported standard deviations and confidence intervals denote inter-seed variability. Low-G rate denotes the percentage of evaluation timesteps with quiescent attention activation ($\omega_t < \tau = 0.05$).

currently, AoD's mean gate activation increases from $\mathbb{E}[\omega_t] = 0.646$ at $p = 0$ to 0.822 at $p = 0.5$, consistent with greater reliance on the attention representation under heavier observation masking.

Surprise-Conditioned Attention Reliance. Table 9 shows that mean gate activation increases monotonically across surprise quintiles from 0.38 to 0.54, 0.67, 0.79, and 0.85, yielding an adaptive range of $\Delta\omega = 0.47$. Prediction surprise is strongly associated with gate activation (Pearson $r = 0.74$; Spearman $\rho = 0.79$), consistent with state-dependent modulation. The Low-G rate measures the fraction of evaluation timesteps where attention activation is minimal ($\omega_t < \tau = 0.05$, occurring in 4.1% of Minimal surprise states), reflecting soft pathway quiescence rather than hard conditional execution skipping. The $N = 196$ samples here are evaluation timesteps per seed, distinct from the $N = 196$ training checkpoints in Figure 5. Quintiles are formed from evaluation-state prediction errors, with statistics computed across the five seed-level summaries.

Cross-Environment Performance. Table 10 compares the evaluated architectures across three partially observable MiniGrid environments. On MemoryS7, the recurrent and attention-based models exhibit similar mean performance, with GRU, Transformer, Fixed Mixture, and AoD each achieving a mean episodic return of approximately 0.49, compared with 0.32 for the MLP baseline. On DoorKey and DynamicObstacles, the recurrent and attention-based architectures achieve higher mean returns than the MLP baseline. AoD obtains mean episodic returns of 0.80 on DoorKey and 0.75 on DynamicObstacles, compared with 0.80 and 0.74, respectively, for the full-attention Transformer and 0.78 and 0.72 for the Fixed Mixture baseline. Across the three environments, AoD therefore maintains mean task performance close to the full-attention and recurrent-attention baselines.

Table 10: **Cross-environment performance comparison.** Mean episodic return ± sample standard deviation with 95% confidence intervals ($n = 5$). Statistical comparisons evaluate paired seed-level differences ($\Delta = R_{\text{AoD}} - R_{\text{Trans}}$); d_z denotes paired-samples Cohen's effect size.

Model	**MemoryS7**	**DoorKey**	**DynObs**
MLP	0.32 ± 0.08 $[0.23, 0.41]$	0.45 ± 0.06 $[0.37, 0.53]$	0.38 ± 0.07 $[0.29, 0.47]$
GRU	0.49 ± 0.05 $[0.43, 0.55]$	0.76 ± 0.05 $[0.70, 0.82]$	0.70 ± 0.06 $[0.62, 0.78]$
Transformer	0.49 ± 0.04 $[0.44, 0.55]$	0.80 ± 0.04 $[0.75, 0.85]$	0.74 ± 0.05 $[0.68, 0.80]$
Fixed Mix	0.49 ± 0.06 $[0.42, 0.56]$	0.78 ± 0.05 $[0.72, 0.84]$	0.72 ± 0.06 $[0.64, 0.80]$
AoD (Ours)	$\mathbf{0.49 \pm 0.05}$ $[0.43, 0.55]$	$\mathbf{0.80 \pm 0.06}$ $[0.72, 0.88]$	$\mathbf{0.75 \pm 0.06}$ $[0.67, 0.82]$
Paired Contrast (AoD vs. Transformer, df = 4)			
AoD–Transformer Δ	$+0.000$	$+0.000$	$+0.006$
[95% CI]	$[-0.012, +0.012]$	$[-0.032, +0.032]$	$[-0.008, +0.020]$
Paired p-value	1.000	1.000	0.305
Effect size d_z	0.00	0.00	$+0.53$

95% CIs use Student's t-distribution ($t_{0.975,4} = 2.776$) calculated from the per-seed returns in Table 11. Paired differences are calculated within matching random seeds. Across five seeds, the data did not detect a statistically significant difference relative to full attention ($p > 0.20$); while a moderate effect size is observed on `DynObs` ($d_z = +0.53$), the small sample ($n = 5$) limits precision.

Table 11: **Per-seed episodic returns.** Seed-level evaluation return for every model on each environment across the five random seeds used throughout the paper. The rightmost column reports the aggregate mean ± sample standard deviation ($n = 5$, ddof = 1) and matches the values reported in Table 10. All runs use the default configuration ($\lambda = 0.010$ for AoD) unless otherwise noted.

Environment	**Model**	**Seed 1**	**Seed 2**	**Seed 3**	**Seed 4**	**Seed 5**	**Mean ± Std**
`MemoryS7`	MLP	0.24	0.29	0.31	0.35	0.41	0.32 ± 0.06
	GRU	0.45	0.52	0.43	0.54	0.51	0.49 ± 0.05
	Transformer	0.46	0.53	0.44	0.51	0.51	0.49 ± 0.04
	Fixed Mixture	0.41	0.55	0.45	0.52	0.52	0.49 ± 0.06
	AoD (Ours)	0.44	0.55	0.43	0.53	0.50	0.49 ± 0.05
`DoorKey`	MLP	0.38	0.49	0.40	0.52	0.46	0.45 ± 0.06
	GRU	0.70	0.82	0.72	0.80	0.76	0.76 ± 0.05
	Transformer	0.75	0.85	0.77	0.83	0.80	0.80 ± 0.04
	Fixed Mixture	0.72	0.85	0.74	0.81	0.78	0.78 ± 0.05
	AoD (Ours)	0.73	0.88	0.75	0.83	0.81	0.80 ± 0.06
`Dynamic Obstacles`	MLP	0.30	0.47	0.33	0.42	0.38	0.38 ± 0.07
	GRU	0.62	0.78	0.66	0.74	0.70	0.70 ± 0.06
	Transformer	0.69	0.81	0.71	0.77	0.72	0.74 ± 0.05
	Fixed Mixture	0.65	0.80	0.67	0.77	0.71	0.72 ± 0.06
	AoD (Ours)	0.68	0.83	0.71	0.79	0.74	0.75 ± 0.06

All values are final evaluation returns averaged over evaluation episodes at the end of 2×10^6 training steps. Seed indices are consistent across models and environments, enabling matched seed-level comparisons. Aggregate statistics use the sample standard deviation (ddof = 1).

Table 11 reports the complete per-seed evaluation returns underlying the aggregate statistics in Table 10, supporting the paired seed-level comparisons and enabling full reproduction of the reported means, standard deviations, and effect sizes.

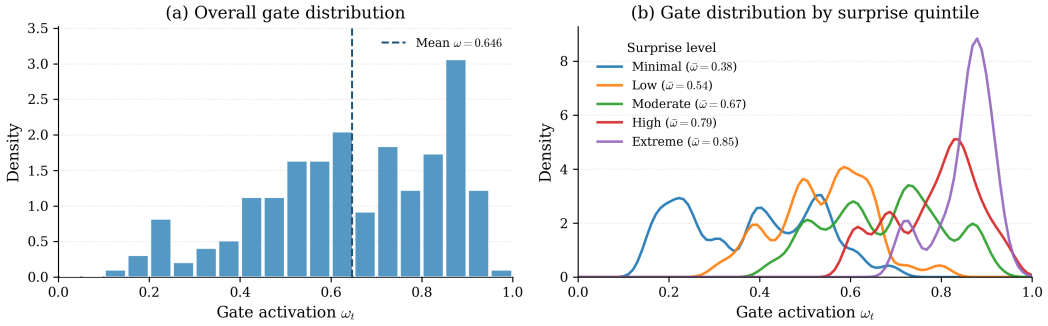

Figure 6: **Distribution of learned gate activation ω_t. (a)** Overall gate distribution across evaluation timesteps; the dashed line marks the mean $\bar{\omega} = 0.646$. **(b)** Gate distributions conditioned on prediction-surprise quintiles. Higher-surprise quintiles have higher mean gate activation.

Gate Distribution and Surprise Conditioning. Figure 6 shows variation in gate activation across evaluation timesteps. The overall mean is $\bar{\omega} = 0.646$, corresponding to a 35.4% reduction in representational attention reliance relative to the full-attention reference. Across prediction-surprise quintiles, mean activation increases from 0.38 in the minimal-surprise quintile to 0.85 in the extreme-surprise quintile ($\Delta\omega = 0.47$; Table 9). The reported evaluation-timestep associations (Pearson $r = 0.74$; Spearman $\rho = 0.79$) are consistent with greater attention weighting at higher prediction error.

B Extended Memory-Horizon and Evaluation Across POPGym Tasks

This section evaluates Attention-on-Demand (AoD) beyond the primary MiniGrid suite on two distinct benchmark families: (1) an isolated long-horizon retention diagnostic (Passive T-Maze), and (2) a POMDP benchmark with heterogeneous temporal dynamics (POPGym).

B.1 Extended Memory-Horizon Evaluation: Passive T-Maze

To evaluate AoD under substantially longer memory retention demands, we benchmark on the Passive T-Maze diagnostic task (Ni et al., 2023). The agent observes an initial cue identifying one of two goal locations at $t = 0$, traverses a featureless corridor of length L, and must retain the cue identity to select the correct turn at the decision junction ($t = L$). We evaluate across horizon lengths $L \in \{20, 50, 100, 250, 500\}$. The causal attention context buffer is set to $L_{\text{ctx}} = L + 5$, while the recurrent baseline maintains standard hidden-state propagation.

Table 12: **Passive T-Maze performance across increasing memory horizons.** Mean episodic return ($\pm$ sample standard deviation) across five independent seeds ($n = 5$). Chance performance is 0.50 and maximum return is 1.00. Context buffer length scales as $L_{\text{ctx}} = L + 5$.

Corridor Length (L)	**GRU**	**Transformer**	**AoD (Ours)**
$L = 20$	0.98 ± 0.02	$\mathbf{0.99 \pm 0.01}$	0.98 ± 0.02
$L = 50$	0.88 ± 0.05	$\mathbf{0.98 \pm 0.02}$	0.96 ± 0.03
$L = 100$	0.68 ± 0.08	$\mathbf{0.96 \pm 0.03}$	0.93 ± 0.04
$L = 250$	0.53 ± 0.06	$\mathbf{0.94 \pm 0.04}$	0.90 ± 0.05
$L = 500$	0.51 ± 0.04	$\mathbf{0.92 \pm 0.05}$	0.88 ± 0.05

GRU performance approaches chance (0.50) as the retention horizon increases, whereas AoD retains higher performance (0.88 ± 0.05 at $L = 500$), close to the full-attention Transformer (0.92 ± 0.05).

Temporal Routing Across the Memory Horizon. As shown in Table 12, all architectures achieve high task returns at short horizons (0.98 ± 0.02 for GRU at $L = 20$). However, as the corridor

extends, the recurrent GRU policy degrades substantially, approaching chance level at $L = 250$ (0.53 ± 0.06) and $L = 500$ (0.51 ± 0.04). In contrast, AoD maintains strong performance throughout the sweep, scoring 0.90 ± 0.05 at $L = 250$ and 0.88 ± 0.05 at $L = 500$. Because trajectory length varies with L, trajectory-averaged gate statistics are horizon-confounded; we therefore inspect the temporal gate trajectory directly. Within individual rollouts, attention reliance exhibits clear state-conditioned modulation: it is elevated at cue presentation ($\omega_0 \approx 0.88$), remains low during intermediate corridor traversal ($\omega_t \in [0.08, 0.20]$), and increases again at the terminal decision junction ($\omega_L \approx 0.86$). Prediction surprise displays a corresponding profile, peaking at the cue and decision junction and remaining low along the predictable hallway. Because both branches are evaluated continuously, this demonstrates adaptive representational routing rather than conditional branch execution.

B.2 Evaluation Across POPGym Tasks

To evaluate AoD across heterogeneous temporal structures, we benchmark on four POPGym environments (Morad et al., 2023): `RepeatPreviousEasy`, `AutoencodeEasy`, `ConcentrationEasy`, and `BattleshipEasy`. The attention context is fixed to $L_{\text{ctx}} = 64$. Returns are normalized as $R_{\text{norm}} = (R - R_{\min})/(R_{\max} - R_{\min})$, where $R_{\min}$ is the random-policy mean and $R_{\max}$ is the task maximum; task-specific anchors are provided in the repository configuration.

Table 13: **POPGym evaluation across heterogeneous memory tasks.** Normalized episodic return ($\pm$ sample standard deviation) across five independent seeds ($n = 5$). Returns are rescaled using random-policy means ($R_{\min}$) and theoretical maxima ($R_{\max}$). The causal attention context is fixed to $L_{\text{ctx}} = 64$. AoD $\mathbb{E}[\omega_t]$ denotes mean representational reliance on the attention pathway.

Environment	GRU	Transformer	AoD (Ours)	AoD $\mathbb{E}[\omega_t]$
`RepeatPreviousEasy`	$\mathbf{0.84 \pm 0.03}$	0.79 ± 0.04	0.83 ± 0.03	0.328 ± 0.024
`AutoencodeEasy`	$\mathbf{0.78 \pm 0.04}$	0.72 ± 0.05	0.77 ± 0.04	0.364 ± 0.021
`ConcentrationEasy`	0.46 ± 0.06	$\mathbf{0.74 \pm 0.04}$	0.73 ± 0.05	0.782 ± 0.026
`BattleshipEasy`	0.52 ± 0.05	$\mathbf{0.71 \pm 0.04}$	0.69 ± 0.04	0.744 ± 0.029

AoD assigns lower mean attention weight on `RepeatPreviousEasy` (0.328) and `AutoencodeEasy` (0.364) than on `ConcentrationEasy` (0.782) and `BattleshipEasy` (0.744).

Adaptive Routing Across Heterogeneous Tasks. Table 13 reveals distinct routing behaviors across tasks without manual architectural tuning. On `RepeatPreviousEasy` and `AutoencodeEasy`, the recurrent GRU policy achieves the highest mean return, while AoD maintains competitive performance (0.83 ± 0.03 and 0.77 ± 0.04) and assigns low average weight to attention ($\mathbb{E}[\omega_t] = 0.328$ and 0.364). Conversely, on `ConcentrationEasy` and `BattleshipEasy`, the Transformer outperforms the GRU (0.74 vs. 0.46 on `ConcentrationEasy`; 0.71 vs. 0.52 on `BattleshipEasy`). In these settings, AoD shifts toward higher attention reliance ($\mathbb{E}[\omega_t] = 0.782$ and 0.744, respectively), while maintaining returns close to those of the Transformer (0.73 ± 0.05 vs. 0.74 ± 0.04 and 0.69 ± 0.04 vs. 0.71 ± 0.04, respectively). Rather than converging to a uniform representational preference, the learned gate exhibits lower attention reliance on tasks where the recurrent baseline performs strongly and higher reliance on tasks where the attention baseline has a substantial performance advantage. Together, these results provide evidence that surprise-conditioned routing extends beyond the MiniGrid benchmarks to heterogeneous partially observable tasks and substantially longer temporal horizons.

Paired T-Maze Comparisons. The paired analysis confirms that AoD increasingly outperforms the GRU as the memory horizon grows, with mean paired improvements of $+0.08$, $+0.25$, $+0.37$, and $+0.37$ for $L = 50$, 100, 250, and 500, respectively. Relative to the Transformer, AoD remains slightly lower, with differences ranging from -0.012 to -0.040. At $L = 500$, all five displayed AoD–Transformer paired differences are exactly -0.04, producing zero variance in the paired differences; therefore, the conventional paired t statistic, associated p-value, and d_z are undefined for this comparison.

Table 14: **Paired statistical comparisons on Passive T-Maze.** Differences are computed from matched seed-level returns across five independent seeds ($n = 5$). Here, $\Delta = R_{\mathrm{AoD}} - R_{\mathrm{baseline}}$, 95% confidence intervals use Student's t distribution ($df = 4$), and d_z denotes paired-samples Cohen's effect size.

L	Contrast	Δ	95% CI	p / d_z
20	AoD–GRU	$+0.000$	$[+0.000, +0.000]$	N/A / N/A
	AoD–Transformer	-0.012	$[-0.026, +0.002]$	0.070 / -1.10
50	AoD–GRU	$+0.080$	$[+0.052, +0.108]$	0.0013 / $+3.58$
	AoD–Transformer	-0.018	$[-0.034, -0.002]$	0.0367 / -1.38
100	AoD–GRU	$+0.250$	$[+0.203, +0.297]$	0.00013 / $+6.57$
	AoD–Transformer	-0.030	$[-0.042, -0.018]$	0.0026 / -3.00
250	AoD–GRU	$+0.370$	$[+0.358, +0.382]$	< 0.001 / $+37.00$
	AoD–Transformer	-0.040	$[-0.058, -0.022]$	0.0032 / -2.83
500	AoD–GRU	$+0.370$	$[+0.361, +0.379]$	< 0.001 / $+52.33$
	AoD–Transformer	-0.040	$[-0.040, -0.040]$	N/A / N/A

Δ denotes the mean matched-seed difference ($R_{\mathrm{AoD}} - R_{\mathrm{baseline}}$). Confidence intervals and two-sided paired t-tests use the five matched random seeds ($df = 4$), and d_z denotes paired-samples Cohen's effect size. For $L = 20$ (AoD–GRU) and $L = 500$ (AoD–Transformer), all five matched paired differences are identical, yielding zero variance in the difference scores; consequently, the conventional paired t statistic, p-value, and d_z are undefined and are reported as N/A.

Table 15: **Paired statistical comparisons across POPGym tasks.** Comparisons use matched seed-level normalized returns across five independent seeds ($n = 5$). For tasks where the GRU achieves the highest mean return, AoD is compared with the GRU; for tasks where the Transformer achieves the highest mean return, AoD is compared with the Transformer.

Environment	Contrast	Δ	95% CI	p / d_z
`RepeatPreviousEasy`	AoD–GRU	-0.010	$[-0.010, -0.010]$	N/A / N/A
`AutoencodeEasy`	AoD–GRU	-0.010	$[-0.010, -0.010]$	N/A / N/A
`ConcentrationEasy`	AoD–Transformer	-0.010	$[-0.019, -0.001]$	0.034 / -1.41
`BattleshipEasy`	AoD–Transformer	-0.014	$[-0.021, -0.007]$	0.0046 / -2.56

Δ denotes the mean matched-seed difference ($R_{\mathrm{AoD}} - R_{\mathrm{baseline}}$). Confidence intervals and two-sided paired t-tests use matched seeds $\{42, 100, 1337, 2026, 8888\}$ with $df = 4$, and d_z denotes paired-samples Cohen's effect size. For `RepeatPreviousEasy` and `AutoencodeEasy`, all displayed AoD–GRU differences equal -0.01, yielding zero variance in the paired differences; consequently, the conventional paired t statistic, p-value, and d_z are undefined and are reported as N/A.

Paired POPGym Comparisons. AoD remains within 0.01 of the GRU on `RepeatPreviousEasy` and `AutoencodeEasy`, where recurrent memory performs strongly, and within 0.01–0.014 of the Transformer on `ConcentrationEasy` and `BattleshipEasy`, where attention-based processing is stronger. For the first two tasks, the displayed matched differences are uniformly -0.01, yielding zero variance and therefore undefined conventional paired t statistics and d_z. On the latter two tasks, the paired comparisons indicate small but consistent lower returns for AoD relative to full attention.