Learning to Explain: Towards Human-Aligned Explainability in Deep Reinforcement Learning via Attention Guidance

Bokai Ji[1] , **Guangxia Li**[†1] , **Yulong Shen**[1] , **Gang Xiao**[2]
[1]Xidian University
[2]National Key Laboratory for Complex Systems Simulation, China
jibokai@stu.xidian.edu.cn, gxli@xidian.edu.cn, ylshen@mail.xidian.edu.cn, searchware@qq.com

Abstract

Recent advances in explainable deep reinforcement learning (DRL) have provided insights into the reasoning behind decisions made by DRL agents. However, existing methods often overlook the subjective nature of explanations and fail to consider human cognitive styles and preferences. Such ignorance tends to reduce the interpretability and relevance of the generated explanations from a human evaluator's perspective. To address this issue, we introduce human cognition into the explaining procedure by integrating DRL with attention guidance in a novel manner. The proposed concept proximal policy optimization (Concept-PPO) learns to generate human-aligned explanations by jointly optimizing the DRL performance and the discrepancy between generated explanations and human annotations. Its key component is a specially designed spatial concept transformer that can enhance explaining efficiency by premasking decision-irrelevant information. Experiments on the ATARI benchmark demonstrate that Concept-PPO achieves better policies than its black-box counterparts, and user studies confirm its superiority in generating human-aligned explanations compared to existing explainable DRL methods.

1 Introduction

A critical impediment to the effective application of deep reinforcement learning (DRL) methods in real-world decision-making scenarios is that they infer in a latent space spawned by a series of neural layers rather than in a manner that is explicit and understandable to humans. It is difficult for practitioners to analyze the connections between input observations and outcome actions of DRL agents. This nontransparency and unpredictability hinders most DRL methods from real-world applications, particularly for those requiring cautious decision and verifiability.

Efforts have been made to achieve explainable reinforcement learning (XRL). Saliency-based XRL methods [Simonyan *et al.*, 2014; Mott *et al.*, 2019; Atrey *et al.*, 2020] explain policies by highlighting critical regions for decision-making in visual inputs. Because salient pixels are irrelevant to high-level concepts that human explainers depend on, saliency-based explanations are often hard to interpret from a human perspective. To explain with high-level concepts inherently, relational learning is employed to identify both the regions crucial to a DRL agent's decisions and the high-level concepts these regions relate to [Zambaldi *et al.*, 2019; Jiang and Luo, 2019; Karia and Srivastava, 2022]. The explanations derived, despite in a human-understandable form, are often impaired by misinterpreted concepts, as there are no exemplary supervisions to guide the generation of human mind compatible explanations.

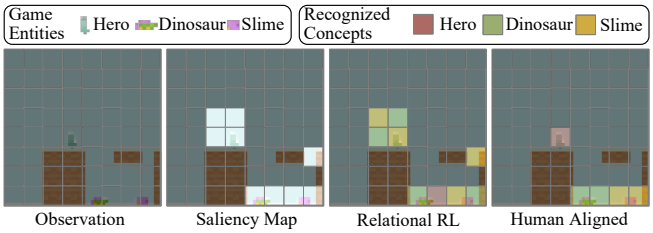


Figure 1: A frame of the CoinRun game from OpenAI Procgen benchmark showing the superiority of aligning explanations with human cognition. Light-color layers are used to indicate important patches that are recognized by agents. Three colors in the legend are used to represent the recognized concepts.

Figure 1 shows the aforementioned flaws of saliency-based RL and relational RL, along with the ideal result obtained when aligning explanations with the human mind. To be explained is a frame of the CoinRun game provided by OpenAI's Procgen benchmark [Cobbe *et al.*, 2020]. There are three visual game entities (see the left in the legend), and three corresponding predefined concepts, namely *Hero*, *Dinosaur*, and *Slime*, that need to be patch-wisely assigned to entities. As human players always focus on visual entities when making decisions, it is expected that an explainer will rightly associate predefined concepts with entities and make good use of them to generate explanations. It can be seen that the saliency map highlights important patches (light gray) that affect the agent's decisions but cannot associate these patches with concepts. The explanation given by relational RL, although based on concepts (it fills patches using the three col-

† Corresponding author.

ors in the legend), deviates from human prior knowledge (the fill-colors are scattered and mismatched with game entities). In contrast, the human-aligned explanation fills patches containing game entities with the right color and right scale, suggesting that the recognized concepts are compliant with the human mind. The failure of saliency-based RL and relational RL indicates that it is infeasible to learn human-aligned explanations of decoupled representations without human supervision and useful inductive bias [Locatello *et al.*, 2019].

Inspired by concept transformer (CT) [Rigotti *et al.*, 2022], a recent method that generates human-aligned explanations through attention guidance for image classification, we aim to obtain human-aligned explanations for XRL. CT is trained on both labeled data and human-annotated explanations by jointly minimizing classification loss and explanation loss. Such a scheme, however, cannot be directly applied to DRL. This is because unlike learning from prepared training examples as supervised learning does, a DRL agent samples data from the environment on the fly according to the policy that is being learned. As the agent's sample generation is outside the annotator's control, it is difficult to selectively annotate explanations. Moreover, manual annotation for every rollout is infeasible because successive DRL frames only differ slightly, making consecutive annotation repetitive and inefficient. Therefore, there is a dearth of supervising information on explanations during the RL process.

To overcome this challenge, we propose to pretrain an attention guider (AG) on a collection of game screens and their corresponding human-annotated explanations. During XRL training, the AG takes the game screen as input and generates explanations that serve as ground truths to supervise the explaining process of the XRL agent. To focus on the few critical entities in observation for decision-making, we improve the CT by breaking the explaining process into two stages: masking patches that are irrelevant for decision-making and associating unmasked patches with relevant concepts, resulting in a novel spatial concept transformer (SCT). Because irrelevant patches are masked out before computing the explanation loss, the SCT avoids the dispersion of errors among all concepts and patches. Unlike CT, which uses a single cross-attention module that struggles with scaling, SCT facilitates the use of more powerful networks for patch masking while preserving the fidelity of explanation (i.e., the degree to which the explanation reflects the decision and aim) [Rigotti *et al.*, 2022]. Theoretical analyses of the fidelity of SCTs explanations are provided in Section 3.6.

Integrating AG and SCT with proximal policy optimization (PPO) [Schulman *et al.*, 2017] results in a novel XRL method, termed concept PPO (Concept-PPO). Experimental evaluation on the ATARI benchmark [Bellemare *et al.*, 2013] demonstrates that Concept-PPO achieved competitive or superior returns compared with those of vanilla black-box PPO. Moreover, a user study involving several saliency-based and relational XRL methods shows that Concept-PPO outperforms baselines both in terms of objective alignment metrics and subjective user preferences. In summary, this study serves as a meaningful attempt at empowering DRL with practical explainability. Our contributions are threefold as follows:

1. We are the first to achieve human-aligned explainability in XRL via attention guidance. This is achieved by a carefully designed pretraining strategy for AG that alleviates the burden of annotating successive DRL frames.
2. We propose a two-stage explanation framework, the SCT, to facilitate supervision of XRL models. We theoretically demonstrate that it can preserve the fidelity of the generated explanations.
3. We provide a concrete algorithm to learn human-aligned explanations for XRL. The superiority of our method was confirmed by a well-designed user study, which provides references for subsequent XRL research.

2 Related Works

To make the decision-making processes of DRL agents interpretable to humans, a bunch of methods [Simonyan *et al.*, 2014; Mott *et al.*, 2019; Atrey *et al.*, 2020; Bertoin *et al.*, 2022; Beechey *et al.*, 2023] turn to computer vision techniques, such as saliency maps and attention maps [Chen *et al.*, 2023; Cornia *et al.*, 2018; Judd *et al.*, 2009], to highlight critical regions in the observation that influence agents' decisions. While useful for visualizing attention, these methods failed to associate these regions with meaningful high-level concepts such as entities and relations, which the human decision-making process relies on, thus limiting the sense and explainability from a human evaluator's perspective.

Relational RL methods [Zambaldi *et al.*, 2019; Jiang and Luo, 2019; Karia and Srivastava, 2022] explain agent policies using high-level concepts and relationships. Although their explanations are structured and semantic, these explanations often misalign with human reasoning owing to two issues. They lack explicit supervision for learning the semantics of predefined high-level concepts, which hinders their ability to generate intuitive explanations. Moreover, in methods that attempt to automatically learn high-level concepts, resulting concepts often have coupled or ambiguous semantics, further diminishing the explainability of explanations.

Our study addresses these limitations by introducing a framework that enables XRL agents to explain using predefined high-level concepts. This allows agents to produce human-aligned explanations that are intuitive, semantically grounded, and practical for real-world applications.

3 Methodology

3.1 Framework

The SCT architecture, as illustrated in Figure 2, is composed of four components: Vision Encoder, Patch Masking Module, Concept Query Module, and Attention Guider. The input is a sequence of frames in the form of RGB images with height H and width W, each of which is referred to as an image observation $O \in \mathbb{R}^{H\times W\times 3}$. A set of learnable concept embeddings $E_c = \{e_c^i \in \mathbb{R}^{d_c}, i \in 1, \ldots, n_c\}$ are also provided, where d_c and n_c denote the embedding length and number of high-level concepts, respectively. The output is an explanation $A_{expl} \in \mathbb{R}^{n_p\times n_c}$ corresponding to the agent's decision-making process in the form of cross-attention scores [Vaswani *et al.*, 2017], and a stochastic policy

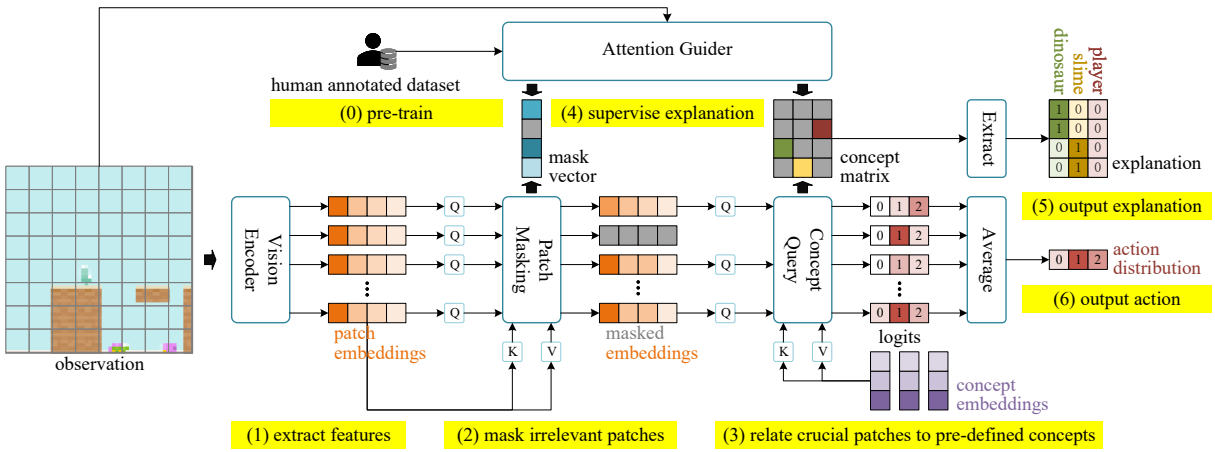


Figure 2: Anatomy of the proposed Spatial Concept Transformer for human-aligned explainability in reinforcement learning.

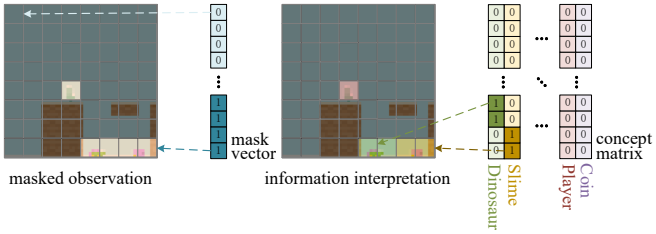


Figure 3: An illustration of the mask vector that serves as a binary indicator delineating the specific patches that are instrumental in the agent's decision-making process. Columns in the concept matrix represent high-level concepts while rows for patches. For example, the green highlighted patches in the information interpretation image are explained as the concept *Dinosaur*, due to a confidence value of 1 at their intersection with the concept matrix.

$Pr(\cdot|O) \in \mathbb{R}^{n_a}$, where n_p is the length of the embedding sequence encoded by the Vision Encoder from O, and n_a is the size of the action space defined by the task.

There is a preparatory step to train an Attention Guider with a manually annotated dataset of human-desired explanations (step (0) in Figure 2). During XRL training, taking an input game frame as an observation, the SCT (1) extracts features into a set of patch embeddings via the Vision Encoder, (2) masks out patches irrelevant to the XRL agent's decision-making process via the Patch Masking Module, (3) generates a concept matrix containing explanations by relating remaining crucial patches to predefined concepts via the Concept Query Module, (4) supervises the generated mask and the concept matrix using the pretrained Attention Guider, and finally (5) outputs explanations by reshaping the concept matrix, and likewise, (6) outputs an action distribution for downstream RL. The key modules are detailed as follows.

3.2 Vision Encoder

The Vision Encoder encodes the RGB observation O into a sequence of patch embeddings $E = \{e_i \in \mathbb{R}^{d_m} | i = 1, \ldots, n_p\}$ to facilitate subsequent patch-wise operations; here, each embedding in E corresponds to a patch in O; d_m is the dimension of the patch embeddings, and n_p is the total number of patches. Specifically, $n_p = (H \times W)/(s_p \times s_p)$, where s_p denotes the patch size. For example, given an input image of shape $210 \times 160 \times 3$, a possible output of the Vision Encoder is $\{e_i \in \mathbb{R}^{128} | i = 1, \ldots, 336\}$. This implies that O is divided into 21×16 patches of 10×10, each of which is encoded into a vector of length 128.

A specially designed convolutional neural network (CNN) is employed to ensure that the Vision Encoder's output matches our desired patch division. The first n convolutional layers are configured with a kernel size of 3, stride of 1, and padding of 1 to capture low-level features while maintaining the shape of the input. The final convolutional layer is configured with a kernel size and stride of both s_p, which should be a common divisor of H and W. This results in a feature map $M \in \mathbb{R}^{\frac{H}{s_p} \times \frac{W}{s_p} \times d_m}$, which can be seen as $\frac{H}{s_p} \times \frac{W}{s_p}$ patch embedding of d_m dimensions. Lastly, each patch embedding undergoes a shared affine transformation to form the output E of the Vision Encoder. Such a setup allows us to perform attention operations at various granularities by dividing the image into patches of different sizes, thereby establishing a balance between precision and efficiency: smaller patches offer more precision, while larger patches enhance learning efficiency by reducing n_p.

3.3 Patch Masking

To identify important regions for decision-making, the Patch Masking Module applies self-attention on input concept embeddings E to generate a mask vector $\mathbf{m} \in [0, 1]^{n_p}$. The value of $\mathbf{m}_p$ is an indicator of patch p's contribution to the agent's decision. As shown in Figure 3, $\mathbf{m}_p \to 1$ indicates more inclination for the agent to consider patch p in making decisions, while $\mathbf{m}_p \to 0$ indicates the converse scenario. Embeddings in E with small values in $\mathbf{m}$ are deemed irrelevant and masked to eliminate their influence on the decision-making process.

Specifically, the input E of the Patch Masking Module serves as queries $Q_\Pi^{n_p \times d_m}$, keys $K_\Pi^{n_p \times d_m}$, and values $V_\Pi^{n_p \times d_m}$ when performing self-attention:

$$\alpha_{\Pi p,p'} = \text{softmax}\left(\frac{Q_\Pi K_\Pi^\top}{\sqrt{d_m}}\right)_{p,p'} \quad (1)$$

The obtained attention map is then averaged along patches to derive the mask vector:

$$\mathbf{m} = \frac{1}{n_p}\sum_{p=1}^{n_p} A_{\Pi p}, \quad (2)$$

where $A_\Pi = [\alpha_{\Pi p,p'}] \in \mathbb{R}^{n_p \times n_p}$ and $p, p' \in 1, \ldots, n_p$.

A_Π is then multiplied by V_Π and the output matrix $O_\Pi^{d_m \times n_p}$ to obtain n_p logits:

$$Logit_\Pi = [A_\Pi V_\Pi O_\Pi]. \quad (3)$$

Applying the mask $\mathbf{m}$ to the logit $Logit_\Pi$ by an element-wise multiplication (Hadamard product here) gives the output of Patch Masking Module:

$$Out_\Pi = \mathbf{m} \cdot Logit_\Pi. \quad (4)$$

Such an operation drives logits corresponding to irrelevant patches toward $\mathbf{0}$, ensuring that only information from critical patches is involved in downstream processing.

3.4 Concept Querying

To associate a crucial patch p with a high-level concept c, the Concept Query Module employs a cross-attention mechanism to calculate the attention score α_{pc} of the two. As shown in Figure 3, $\alpha_{pc} \to 1$ indicates a strong association between the patch p and the concept c, while $\alpha_{pc} \to 0$ indicates the opposite. The output logits of the Concept Query Module are transformed into the agent's action distributions through a linear operation, which preserves the explainability of attention scores as suggested by a previous study [Alvarez-Melis and Jaakkola, 2018] and confirmed via our theoretical analysis in Section 3.6.

Specifically, by setting Out_Π as queries $Q_\Lambda^{n_p \times d_m}$ and concept embeddings E_c as keys $K_\Lambda^{n_c \times d_m}$ and values $V_\Lambda^{n_c \times d_m}$, the concept querying is performed as follows:

$$\alpha_{\Lambda p,c} = \text{softmax}(\frac{Q_\Lambda K_\Lambda^\top}{\sqrt{d_m}})_{p,c} \quad (5)$$

The cross-attention scores derived are organized into a concept matrix $A_\Lambda = [\alpha_{\Lambda p,c}] \in \mathbb{R}^{n_p \times n_c}$, which is enforced to align with human cognition by the supervision from Attention Guider, as detailed in Section 3.5. Applying shape transformation on A_Λ gives to the output explanation A_{expl}.

Meanwhile, by mean pooling over the product of A_Λ, V_Λ, and the output matrix $O_\Lambda \in \mathbb{R}^{d_m \times n_a}$, we obtain the action execution probability distribution of the agent:

$$Logit_{\Lambda i} = \frac{1}{n_p}\sum_{p=1}^{n_p}[A_\Lambda V_\Lambda O_\Lambda]_{p,i} \quad (6)$$

where $i \in 1, \ldots, n_a$ and n_a is the size of action space.

The abovementioned linear average operation ensures that the agent's action distribution is strictly in accordance with the concept matrix A_Λ, or, shortly, it takes actions in an explainability-preserving manner. To see the reason, we note that Equation 6 indicates that given an input x, the conditional probability of executing action i is given as follows:

$$Pr(i|x) = \text{softmax}_i\left(\sum_{c=1}^{n_c}\beta_c\gamma_c(x)\right) \quad (7)$$

where β_c is defined as $(\beta_c)_i = [V_\Lambda O_\Lambda]_{c,i}$, and $\gamma_c(x) = \frac{1}{n_p}\sum_{p=1}^{n_p}\alpha_{\Lambda p,c}$ represents the average contribution of concept c to the final decision given the input x. As each β_c denotes the action execution distribution when the agent only takes concept c for decision-making, we can obtain the overall action distribution by weighting β_c with $\gamma_c(x)$ and assembling them all. The weighting term $\gamma_c(x)$, representing the contribution of concept c to decisions, provides faithful explanations of the agent's policy.

3.5 Attention Guider

To align explanation $A_{expl} = [a_{pc}]_{n_p \times n_c}$ with human cognition, we pretrain an Attention Guider with a few annotated game frames and use it to generate ground truth $A_{guide} \in \mathbb{R}^{n_p \times n_c}$ according to the input observation to provide supervision signals for the Concept Query Module and the Patch Masking Module. Specifically, for the Concept Query Module, $A_{guide}^{pc} = 1$ if the agent is expected to consider patch p and associate it with concept c; otherwise, $A_{guide}^{pc} = 0$. For the Patch Masking Module, $\mathbf{m}_{guide}^p = \mathbf{1}\left(\sum_c A_{guide}^{pc}\right)$, where $\mathbf{1}(\cdot)$ represents the indicator function. These two modules are jointly optimized during XRL learning using the following loss:

$$\mathcal{L}_{expl} = \|\mathbf{m} - \mathbf{m}_{guide}\|_2^2 + \lambda\|A_\Lambda^{m-} - A_{guide}^{m-}\|_F^2 \quad (8)$$

where $A_\cdot^{m-}$ denotes the attention scores of unmasked patches, $\|\cdot\|_F$ is the Frobenius norm, and $\lambda \geq 0$ is a parameter balancing the contributions of the patch masking loss and concept query loss. Compared with a previous study that guides the predicted cross-attention map as a whole [Rigotti *et al.*, 2022], our approach ensures that only unmasked patches in a frame contribute to the concept query loss, thereby enhancing the density of effective supervision signals.

To achieve human-aligned explainability **without** sacrificing the performance on DRL tasks, we jointly optimize the explanation loss and RL policy as follows:

$$\mathcal{L} = \mathcal{L}_{rl} + \psi\mathcal{L}_{expl} \quad (9)$$

where $\mathcal{L}_{rl}$ denotes any loss in DRL methods (e.g., sum of the value estimation loss and policy loss in PPO) and $\psi \geq 0$ is a coefficient balancing the contributions of $\mathcal{L}_{expl}$ and $\mathcal{L}_{rl}$.

3.6 Theoretical Analysis

We present theoretical analysis on the effectiveness of Patch Masking Module and the fidelity of explanations given by SCT. First, for the mask vector $\mathbf{m}$, we claim that:

Proposition 1. *The mask vector* $\mathbf{m}$ *can accurately mask patches irrelevant to the decision-making process by causing their cross-attentions with any concept to approach to* 0*. That is, when* $\mathbf{m}_p \to 0$*, it guarantees that* $\alpha_{\Lambda p,c} \to 0$ *for any concept* c*.*

	PPO	Concept-PPO
Pong	$\mathbf{20.60 \pm 0.49}$	20.20 ± 0.75
Seaquest	438.00 ± 9.80	$\mathbf{770.00 \pm 17.89}$
SpaceInvaders	1040.00 ± 24.49	$\mathbf{1232.00 \pm 39.19}$
Enduro	218.00 ± 82.70	$\mathbf{336.00 \pm 89.90}$
Assault	2852.00 ± 886.05	$\mathbf{3415.00 \pm 1167.12}$
BattleZone	17000.00 ± 5357.24	$\mathbf{28200.00 \pm 5004.00}$

Table 1: Episodic returns in 5 test episodes after 5×10^6 time-steps training on ATARI games.

This property ensures that patches considered irrelevant to the decision-making process will not be processed by the following Concept Query Module. Moreover, we show that SCT is guaranteed to be of fidelity for the explanation—a crucial trait that is indispensable for XRL methods. We formalize the fidelity of explanation provided by the SCT as:

Proposition 2. *Any $\gamma_c(x)$ in Equation (7) is a faithful explanation of the stochastic policy. Specifically, by decreasing $\gamma_c(x)$ while keeping $\gamma_{c' \neq c}(x)$ unchanged, the difference between $Logit_\Lambda$ and β_c is monotonically non-decreasing. That is, when keeping $\gamma_{c' \neq c}(x)$ unchanged, $D_{KL}(\beta_c \| Logit_\Lambda)|_{\gamma_c(x)=\Gamma} \leq D_{KL}(\beta_c \| Logit_\Lambda)|_{\gamma_c(x)=\Gamma-\epsilon}$, where $0 \leq \epsilon \leq \Gamma$ and D_{KL} is the Kullback-Leibler (KL) divergence.*

This indicates that as the explanation weight $\gamma_c(x)$ for concept c decreases, the difference between the overall action distribution $\sum_{c=1}^{n_c} \beta_c \gamma_c(x)$ and the action distribution β_c that considers only concept c is monotonically non-decreasing. Therefore, $\gamma(x)$ reflects the contribution of each concept and provides faithful explanation of the agent's policy.

4 Reinforcement Learning Performance

We employed PPO [Schulman *et al.*, 2017] as our training framework. Specifically, we replaced the actor in PPO with the proposed SCT to instantiate an explainable DRL method named Concept-PPO. We demonstrated performance improvement in RL by our Concept-PPO on six games using the ATARI benchmark [Bellemare *et al.*, 2013]. The comparison method was a normal PPO using CNN as the visual head, along with two separate linear layers as the actor and the critic, respectively. The same training procedure was used for both Concept-PPO and normal PPO.

Table 1 presents the episodic returns of Concept-PPO and normal PPO in 5 test episodes after 5×10^6 time-steps of training. One can observe that in the Pong game, the performance of Concept-PPO was comparable to that of normal PPO, while in the other five games, the improvement was significant (20% increase in return on an average). The high returns of Concept-PPO validated our assumption that attention guidance helps DRL agent focus more on task-relevant information and ignore distractors. The suboptimal performance of Concept-PPO in the Pong game may be attributed to the fact that compared with the graphics of the other five games, the graphics of the Pong game are excessively simple (e.g., fewer entities in a frame and fewer ornaments, which are irrelevant to the task); thus, the improvement provided via attention guidance was less than in the Pong game.

In addition to achieving superior RL returns, Concept-PPO also generated human-aligned explanations. To evaluate, we performed a comprehensive user study to evaluate its human-aligned explainability on three of the six ATARI games.

5 User Study on Explainability

This user study aimed to evaluate the explanations of XRL methods and the degree to which they aligned with human cognition. In the evaluation, we paid attention to two factors: 1) how well the XRL agents identified crucial visual regions that facilitated decision-making, and 2) how effectively they related these regions to human-understandable concepts.

5.1 Study Setup

Concept-PPO was tested on three ATARI games (Pong, SpaceInvaders, and BattleZone) against three established XRL approaches, namely Jacobian-based saliency map (JSM) [Zahavy *et al.*, 2016], perturbation-based saliency map (PSM) [Greydanus *et al.*, 2018], and attention-augmented (AA) [Mott *et al.*, 2019]. JSM and PSM highlighted regions that the RL agent focused on but did not relate them to high-level concepts. AA linked focused regions with autonomously discovered concepts using the multi-head attention mechanism. Their explanations for a single game frame are shown in Figure 4. Notably, we merely duplicated AA's explanations according to its site[1]. Since AA employed four attention heads, there were correspondingly four pieces of explanations. More results and videos have been hosted on our GitHub repository[2].

We recruited 30 native English speaking participants from *prolific.co*[3] — a well-known crowd-sourcing platform for online studies. A questionnaire was designed to assess human participants' perceptions of the generated explanations. Participants were warmed up by introducing visual concepts of the game, such as *Player*, *Ball*, *Opponent*, and *Score Board*. By examining the answers to the first question *Q1: "What visual concepts in this game do YOU THINK are most important while playing?"*, we captured participants' subjective perspectives on critical gameplay elements, i.e., a subset of predefined visual concepts. They served as a reference for assessing the alignment between human understanding and the explanations provided by the XRL methods.

We then presented participants with videos that contained explanations generated by XRL methods and asked them several carefully designed questions as follows. For JSM and PSM, the question was *Q2: "Where do you think the RL agent is focusing on?"* Responses could include one or more predefined visual concepts or the option *"Some random and meaningless regions"*, which accounted for the case in which the agent's focus seemed arbitrary.

For AA, participants had to evaluate explanations provided by its four attention heads individually, and answer the question *Q3: "What predefined concept do you think this attention head is focusing on?"* Options were to select predefined concepts, or to choose *Hard to tell* if the regions attended could not be clearly interpreted.

[1]https://sites.google.com/view/s3ta

[2]https://github.com/Bokai-Ji/AG-Policy

[3]https://www.prolific.com

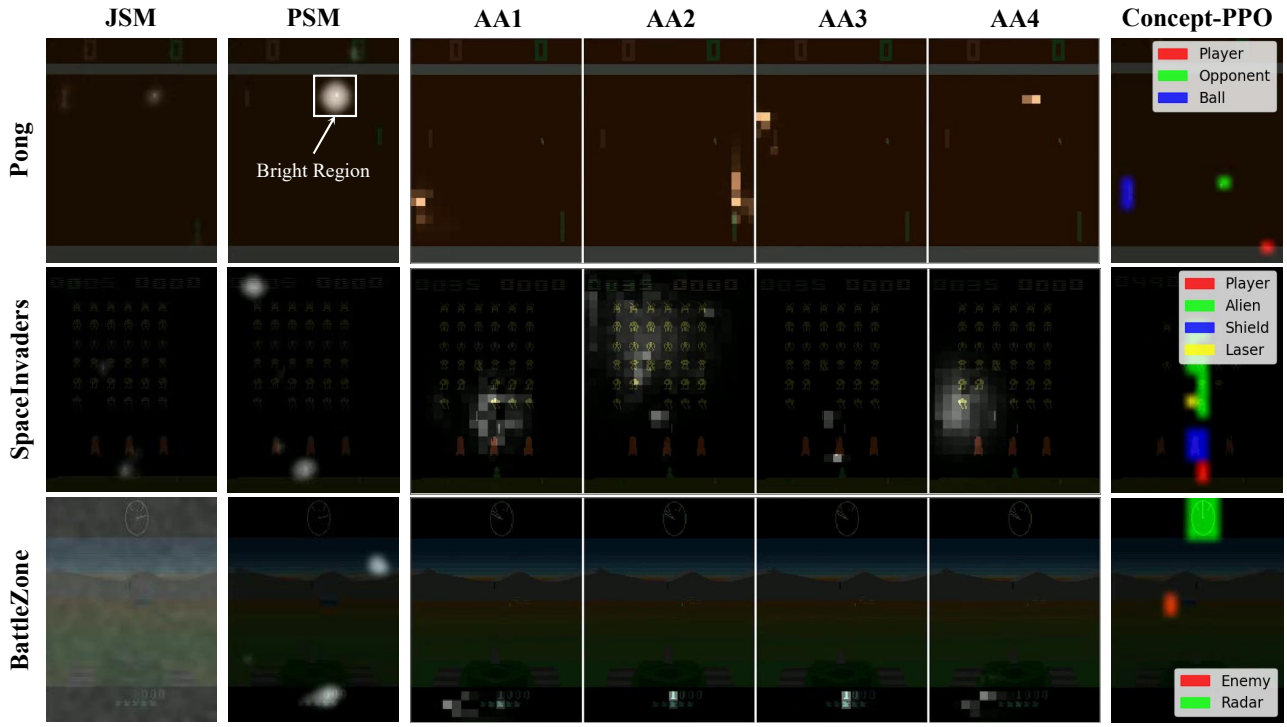


Figure 4: An example of explanations for a single frame of three games (Pong, SpaceInvaders, and BattleZone) given by four XRL methods (JSM, PSM, AA, and Concept-PPO). AA renders its explanations through four attention heads (AA1, AA2, AA3, and AA4). Bright regions are of high attentions, while distinct colors in Concept-PPO represent different high-level concepts that the agent associate these patches with. The label "Bright Region" in PSM was not by the algorithm, but was post added for clarity.

Game	Concept	Human-*Q1*	XRL						
			JSM-*Q2*	PSM-*Q2*	AA-1-*Q3*	AA-2-*Q3*	AA-3-*Q3*	AA-4-*Q3*	Ours-*Q2*	Ours-*Q4*
Pong	Player*	**71.4**	**78.6**	**78.6**	7.1	28.6	14.3	10.7	**78.6**	86.7
	Ball*	**96.4**	**75.0**	**89.3**	21.4	10.7	14.3	28.6	**85.7**	93.3
	Opponent*	**57.1**	**71.4**	**57.1**	**35.7**	14.3	**35.7**	7.1	**67.9**	76.7
	Score Board	10.7	46.4	**67.9**	0.0	3.6	3.6	10.7	7.1	-
	Meaningless	-	17.9	3.6	**35.7**	**42.9**	32.1	**42.9**	3.6	-
SpaceInvaders	Score Board	14.3	21.4	**53.6**	0.0	3.6	3.6	0.0	17.9	-
	Command Ship*	46.4	28.6	25.0	7.1	0.0	7.1	0.0	46.4	55.2
	Alien*	**82.1**	**57.1**	42.9	**46.4**	17.9	3.6	25.0	**78.6**	89.7
	Laser*	**67.9**	**75.0**	**78.6**	3.6	**46.4**	21.4	10.7	**71.4**	79.3
	Shield*	**67.9**	28.6	35.7	0.0	3.6	0.0	7.1	**71.4**	79.3
	Player*	**75.0**	**85.7**	**71.4**	10.7	7.1	**46.4**	7.1	**82.1**	86.2
	Meaningless	-	3.6	10.7	32.1	21.4	17.9	**50.0**	14.3	-
BattleZone	Radar*	**67.9**	44.8	50.0	6.7	3.4	3.3	3.3	**83.3**	93.3
	Cross Hair	**75.0**	27.6	6.7	0.0	10.3	10.0	3.3	20.0	-
	Enemy*	**85.7**	27.6	30.0	3.3	6.9	10.0	3.3	**76.7**	86.7
	Player	50.0	34.5	**63.3**	26.7	**34.5**	36.7	36.7	16.7	-
	Score Board	17.9	24.1	30.0	26.7	17.2	13.3	13.3	16.7	-
	Meaningless	-	44.8	50.0	**36.7**	27.6	**36.7**	**40.0**	6.7	-

Table 2: A numerical summarization of the questionnaire results showing the percentage of participants who say "aye" for every question and option. Each column corresponds to a question dedicated to a particular XRL method (AA has four columns for its four attention heads). And each row represents a visual concept of the game which serves as an answer for the question (an exception is the extra option "Meaningless" which is not a visual concept). For example, the number 74.1 in the column "Human" and the row "Pong-Player" stands for the percentage of participants who selected "Player" as the answer to question *Q1: "What visual concepts ...?"* in the game Pong.

For Concept-PPO, after being shown the concepts that the agent was trained to focus on, along with the corresponding color coding for explanations, participants were asked the same question as *Q2*, and *Q4: "Among concepts that we wish the RL agent to focus on, what concepts are correctly understood by the RL agent (i.e., marked with the correct color in the video)?"* Options to the first question were identical to those of JSM and PSM. And the answer to the second question could include one or more visual concepts that Concept-PPO had been trained to focus on.

Lastly, to assess user's overall preference, we asked the participants a question *Q5: "Which of the four types of explanation do you think best satisfies the criteria?"*, and set the criteria as *"(1) It focuses on reasonable regions that align with your opinion; (2) It effectively relates focused regions to predefined concepts"*. *Q5* provided insights into participants' subjective evaluation of the utility and clarity of explanations.

Through the evaluation of responses to these questions, we

could obtain a measure of how well Concept-PPO provided explanations that were more human-understandable and more aligned to the human mind, and an insight into the subjective preferences of participants.

5.2 Results and Analysis

Table 2 presents the results of the questionnaire by calculating the percentage of participants who responded with "aye" for every question and option.

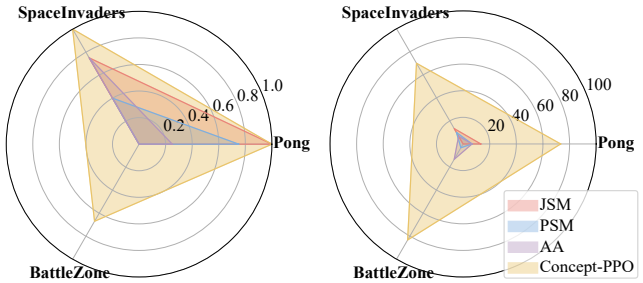
Alignment with Human Attention

We first evaluated how well the XRL method identified regions that human participants regarded as critical for gameplay. Visual concepts selected by $>50\%$ of participants when answering *Q1* formed a human reference set $\mathcal{H}$. Similarly, for JSM, PSM, and Concept-PPO, visual concepts selected by $>50\%$ of participants when answering *Q2* formed the corresponding XRL visual attention set $\mathcal{R}$. Because no visual concept was selected by $>50\%$ of participants when answering *Q3*, in the case of AA, we defined $\mathcal{R}$ as the union of most selected region by its four attention heads. The degree of alignment between XRL's and human's attentions was measured by attention intersection over union (A-IoU): $[(\mathcal{R} \cap \mathcal{H})/(\mathcal{R} \cup \mathcal{H})] \in [0, 1]$. A higher A-IoU value indicated better alignment between the XRL and human cognition.

Figure 5a shows A-IoU values across different XRL methods and games. AA's poor alignment was as per expectation, as it often focused on wide or meaningless regions that lack clear conceptual relevance, as clearly shown in Figure 4. JSM and PSM, as evidenced by their moderate A-IoU scores, tended to be poorly aligned with humans as the underlying graphic became complex. This was because they merely focused on regions with prominent low-level features (e.g., regions with high contrast relative to the background) but not task-relevant regions. In contrast, Concept-PPO achieved the strongest alignment with human cognition, with A-IoU values of 1, 1, and 0.67 on the games Pong, SpaceInvaders, and BattleZone, respectively. The strong human-aligned explainability of Concept-PPO attributed to enforced alignment with human-annotated examples, which made the XRL agent aware of human preferences for policy explanation.

Relating to Human-Understandable Concepts

We next evaluated how well Concept-PPO associated focused regions with human-defined concepts. The last column named "Ours-*Q4*" in Table 2 shows the percentage of participants who agreed that Concept-PPO correctly associated a focused visual region with a human-defined concept when the concept did appear in that region. Only concepts that human annotators considered critical to gameplay were included in this evaluation (marked using "*" in the table). The percentage of participants who thought Concept-PPO had successfully related critical regions with appropriate concepts was moderately high (approximately 82% on an average), highlighting the effectiveness of learning the semantics of concepts and leveraging them in decision-making. Notably, JSM, PSM, and AA could not even perform this evaluation because the former two did not relate visual regions to concepts and the latter was ignorant of predefined concepts. This further



(a) A-IoU (b) User preference

Figure 5: Evaluation results where the radial axis represents the attention intersection over union (A-IoU) scores (left), and the percentage of testers who prefer Concept-PPO's explanations (right).

demonstrated the advantages obtained by learning human-aligned explanations for XRL in the proposed manner.

Subjective Preferences

The question *Q5* listed in Section 5.1 evaluated testers' preference for explanations given by four XRL methods. From Figure 5b showing the preference in percentage, we can see that Concept-PPO was overwhelmingly favored, with $>70\%$ of participants selecting it as the most satisfactory XRL method in games with simpler graphics (Pong and SpaceInvaders) and $>80\%$ in more complex scenarios (BattleZone). In contrast, JSM and PSM were largely criticized for the lack of high-level semantic connections, while AA was deemed unhelpful because of its inability to provide human-understandable explanations.

The aforementioned results indicate that Concept-PPO significantly outperformed existing methods in terms of aligning attention with human cognition and mapping focused regions to predefined concepts. These strengths were reflected in both objective alignment metrics and subjective user preferences, highlighting the importance of incorporating human supervision into the development of XRL explanations.

6 Conclusion

We present an XRL method that incorporates attention guidance to generate high-level human-aligned explanations in the form of attention weights from unstructured visual inputs. By pretraining an Attention Guider, we address the inefficiency of manual annotation in RL tasks. We facilitate the attention guidance procedure by decoupling the guidance of explanation generation into two stages: patch masking and concept querying, resulting in an SCT. Integrating it with PPO yields the Concept-PPO, which demonstrates competitive or superior performance on the ATARI benchmark compared to black-box PPO models while producing explanations that are both precise and human-aligned. Our user study indicates that participants preferred Concept-PPO's explanations considerably more than existing XRL approaches. These results demonstrate the effectiveness of integrating human knowledge to construct human-aligned XRL agents. In future work, we wish to extend our method to provide long-term behavioral explanations, which we believe is more challenging but worthy of further study.

Acknowledgments

This work was supported by the Major Research Plan of the National Natural Science Foundation of China (Grant No. 92267204).

References

[Alvarez-Melis and Jaakkola, 2018] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Neural Information Processing Systems*, 2018.

[Atrey *et al.*, 2020] Akanksha Atrey, Kaleigh Clary, and David D. Jensen. Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning. In *International Conference on Learning Representation*, 2020.

[Beechey *et al.*, 2023] Daniel Beechey, Thomas M. S. Smith, and Özgür Simsek. Explaining reinforcement learning with shapley values. In *International Conference on Machine Learning*, 2023.

[Bellemare *et al.*, 2013] Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *J. Artif. Intell. Res.*, 2013.

[Bertoin *et al.*, 2022] David Bertoin, Adil Zouitine, Mehdi Zouitine, and Emmanuel Rachelson. Look where you look! saliency-guided q-networks for generalization in visual reinforcement learning. In *Neural Information Processing Systems*, 2022.

[Chen *et al.*, 2023] Shi Chen, Ming Jiang, and Qi Zhao. What do deep saliency models learn about visual attention? In *Neural Information Processing Systems*, 2023.

[Cobbe *et al.*, 2020] Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International Conference on Machine Learning*, 2020.

[Cornia *et al.*, 2018] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Trans. Image Process.*, 2018.

[Greydanus *et al.*, 2018] Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. Visualizing and understanding atari agents. In *International Conference on Machine Learning*, 2018.

[Jiang and Luo, 2019] Zhengyao Jiang and Shan Luo. Neural logic reinforcement learning. In *International Conference on Machine Learning*, 2019.

[Judd *et al.*, 2009] Tilke Judd, Krista A. Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision*, 2009.

[Karia and Srivastava, 2022] Rushang Karia and Siddharth Srivastava. Relational abstractions for generalized reinforcement learning on symbolic problems. In *International Joint Conference on Artificial Intelligence*, 2022.

[Locatello *et al.*, 2019] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Reproducibility in Machine Learning, International Conference on Learning Representation Workshop*, 2019.

[Mott *et al.*, 2019] Alexander Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, and Danilo Jimenez Rezende. Towards interpretable reinforcement learning using attention augmented agents. In *Neural Information Processing Systems*, 2019.

[Rigotti *et al.*, 2022] Mattia Rigotti, Christoph Miksovic, Ioana Giurgiu, Thomas Gschwind, and Paolo Scotton. Attention-based interpretability with concept transformers. In *International Conference on Learning Representations*, 2022.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, 2017.

[Simonyan *et al.*, 2014] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations*, 2014.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[Zahavy *et al.*, 2016] Tom Zahavy, Nir Ben-Zrihem, and Shie Mannor. Graying the black box: Understanding dqns. In *International Conference on Machine Learning*, 2016.

[Zambaldi *et al.*, 2019] Vinícius Flores Zambaldi, David Raposo, Adam Santoro, et al. Deep reinforcement learning with relational inductive biases. In *International Conference on Learning Representations*, 2019.