# Survey on Explainable Reinforcement Learning

This survey is part of research focused on making reinforcement learning (RL) algorithms more explainable. RL is a technique used to make decisions in tasks like playing games or controlling robots. For example, AlphaGo, a famous RL-based system, defeated the world champion in the game of Go. Today, RL algorithms are often combined with deep learning to achieve high performance. However, while these algorithms perform well, they are often considered "black boxes" — it's not always clear why they make certain decisions.

In this survey, **we present explanations gained from different methods about 1) Where the RL agent is focusing on and 2) How does the RL agent understand the focused region.**
Specifically, we are using the scenario of playing ATARI video games. In this context, an RL agent is shown frames from the game and must make decisions (such as controlling the game character). We will present different methods that explain the agent's decisions, either by addressing the first question (what the agent considers) or the second question (how the agent understand the information).

**Read this page, and click "Next Page" at the bottom when you are ready to answer the questions (i.e., you know exactly what does explanations from different methods mean).**

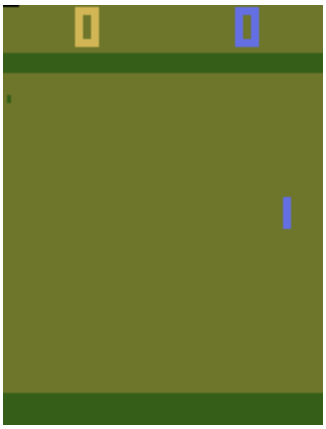
* 表示必填

## 1 Preliminaries

We introduce the **meaning of each type of the explanation** here to get you ready for the questions. **You don't have to understand their names if they are strange to you, it's just for identification purpose.**

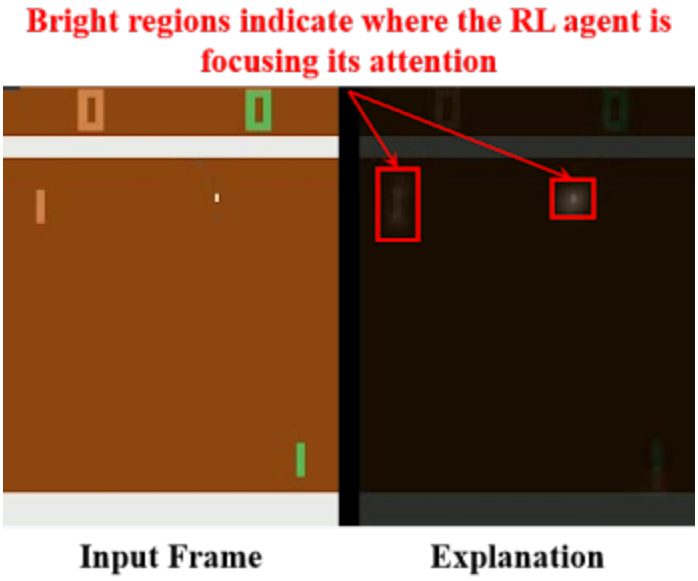
### ATARI Game "Pong"

We take ATARI game Pong as an example to introduce each type of the explanations. Pong is played much like tennis. Each player rallys the ball by moving the paddles on the playfield. We take this simple game as an example to help you understand meaning of the given explanations. We define names of the visual concepts in Pong: 1) ***Score Board:*** Area on the top of the screen that contains numbers; 2) ***Player:*** Green paddle on the right side, controlled by human or RL agents; 3) ***Opponent:*** Orange paddle on the left side controlled by others; 4) ***Ball:*** White spot moving left and right.



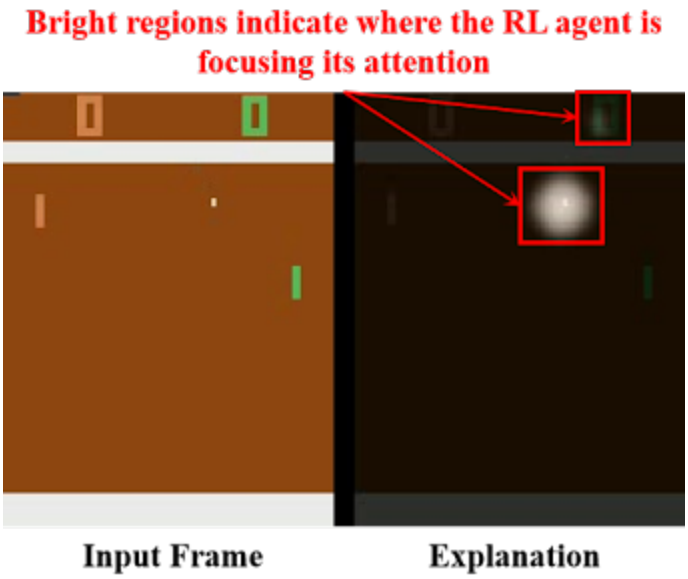
**Explanation-Type 1: Jacobian-based Saliency Map (JSM)**

The right side of the figure below visualizes JSM's explanation in terms of visual attention. Brighter regions indicate where the RL agent is focusing, with the brightness level reflecting the strength of the agent's attention. For instance, the following figure shows that the RL agent focuses on regions we marked out.



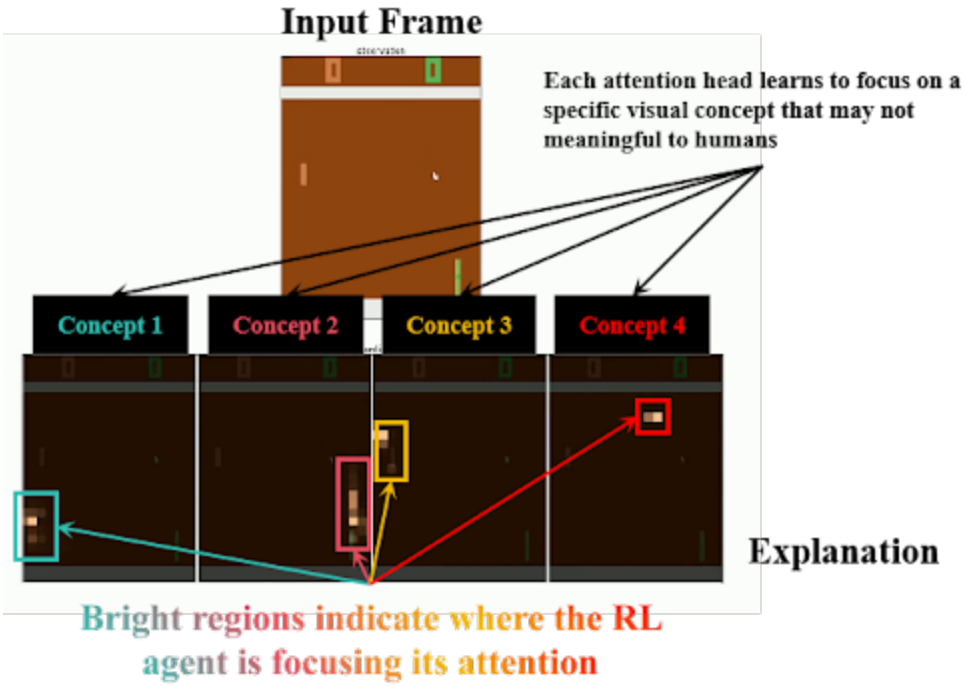
**Explanation-Type 2: Perturbation-based Saliency Map (PSM)**

The right side of the figure below visualizes JSM's explanation in terms of visual attention. Brighter regions indicate where the RL agent is focusing, with the brightness level reflecting the strength of the agent's attention. For instance, the following figure shows that the RL agent focuses on regions we marked out.



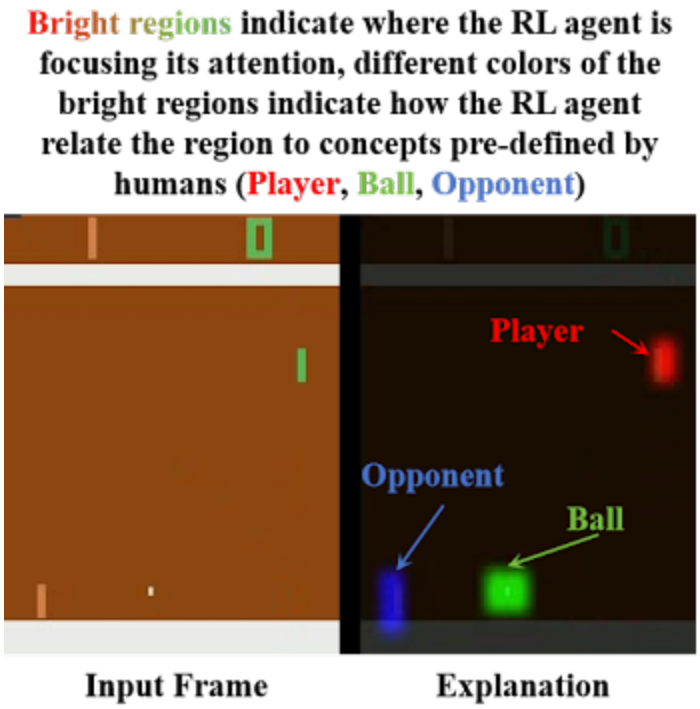
**Explanation-Type 3: Attention Augmented (AA)**

As shown in the figure below, AA's explanation consists of four sub-explanations, each highlighting the agent's visual attention on a specific concept (automatically identified by itself and may not reasonable to humans). In other words, AA demonstrates its understanding of the regions it focuses on by assigning attention to different visual concepts. For instance, in the figure, Concept2 and Concept3 likely represent the agent's attention on the Player and the Opponent, while Concept1 and Concept4 do not seem to correspond to meaningful visual concepts from humans' perspective.



**Explanation-Type 4: Spatial Concept Transformer (SCT)**

Right side of the following figure shows the visualization of SCT's explanation. Bright regions indicate where the RL agent is focusing its attention and different colors of these regions indicate how SCT associate each region with human-defined concepts (e.g., Player, Ball, Opponent). It is reasonable for SCT to not focus on the Score Board, as it does not contribute to improving the control strategy.




**Game 1: Pong**

Now we start the real evaluation with the game Pong, the one we used as a guidance in previous, to warm you up.
**NOTE:** Some videos are long and you don't have to finish it. **Stop watching as long as you are ready for the questions.**

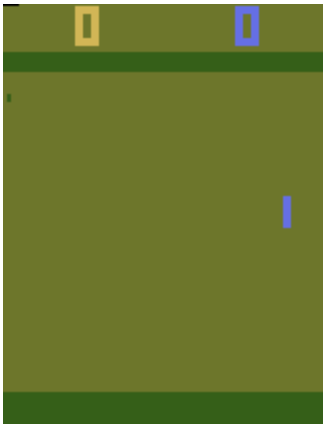
**!!!!!!!!!!!!!!!!!!!!!!  Some questions may have Multiple Selections !!!!!!!!!!!!!!!!!!!!!!!**

**[Multiple Selection]:** multiple choice questions

**[Single Selection]:** the question can have only one answer

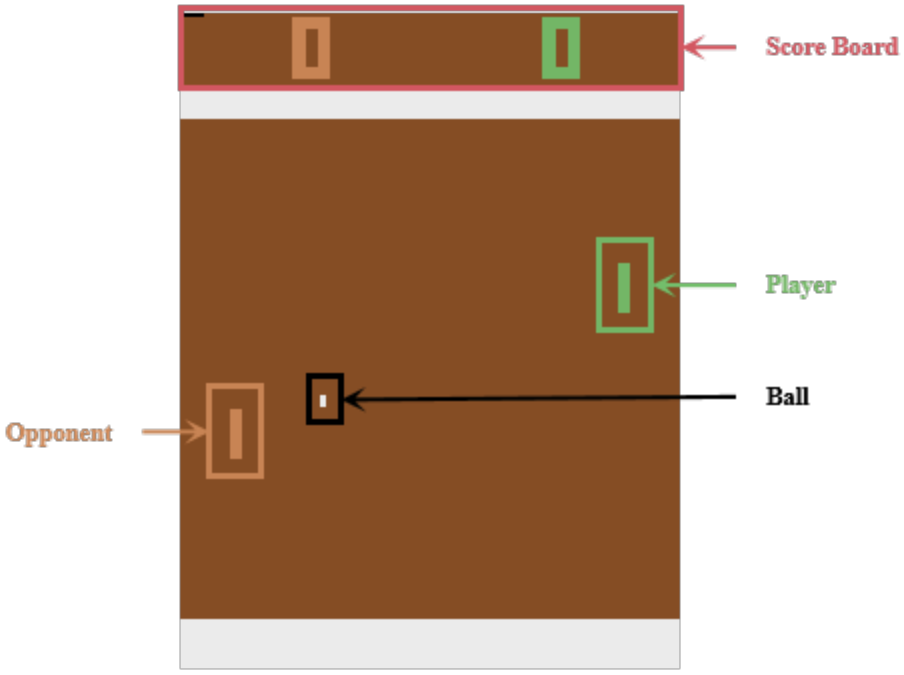
**Description of the game**

We take ATARI game Pong as an example to introduce each type of the explanations. Pong is played much like tennis. Each player rallys the ball by moving the paddles on the playfield. We take this simple game as an example to help you understand meaning of the given explanations.



**Visual Concepts**

***Score Board:*** Area on the top of the screen that contains numbers.
***Player:*** Green paddle on the right side, controlled by human or RL agents.
***Opponent:*** Orange paddle on the left side controlled by others.
***Ball:*** White spot moving left and right.




1。 **[Multiple Selection]** What visual concept in this game do **YOU THINK** are most important while you are playing this game? ( A visual concept is considered crucial if removing it from the screen would hinder your ability to make effective **immediate reactions**.) *

*请选择所有适用项。*

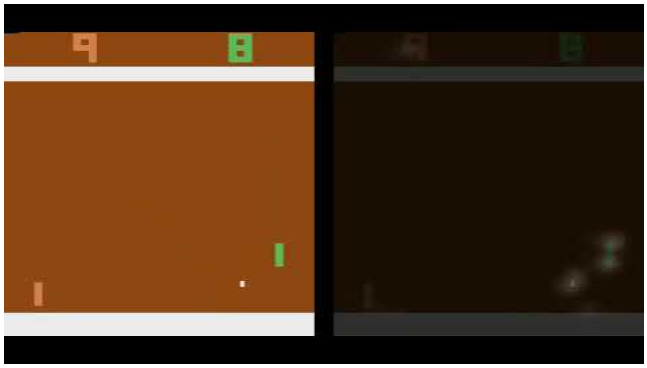
- [ ] Player
- [ ] Ball
- [ ] Opponent
- [ ] Score Board

The RL agent may focus on areas that seem unreasonable to you. However, simply select the regions where the agent is objectively focusing. Please note, the following questions are **NOT** asking where you think the agent should focus, but where it is actually focusing.

**Explanation-Type 1: Jacobian-based Saliency Map (JSM)**

Imagine that the RL agent is sitting in the front of a TV. The left side of the video below shows the TV screen. The right side of the video below visualizes the RL agent's visual attention. Brighter regions indicate where the RL agent is focusing.

**YOU SHOULD:** Read through the questions and choices before watching the video.



http://youtube.com/watch?v=LRsdiWx38QY

2。 **[Multiple Selection]** Based on the right side of the video, select one option for each bright region. *

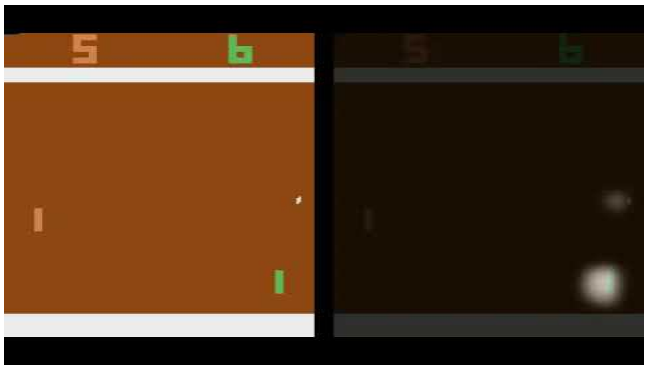
*请选择所有适用项。*

- [ ] Player
- [ ] Ball
- [ ] Opponent
- [ ] Score Board
- [ ] Some random, meaningless regions

**Explanation-Type 2: Perturbation-based Saliency Map (PSM)**

Imagine that the RL agent is sitting in the front of a TV. The left side of the video below shows the TV screen. The right side of the video below visualizes the RL agent's visual attention. Brighter regions indicate where the RL agent is focusing.

**YOU SHOULD:** Read through the questions and choices before watching the video.



3。 **[Multiple Selection]** Based on the right side of the video, select one option for each bright region. *

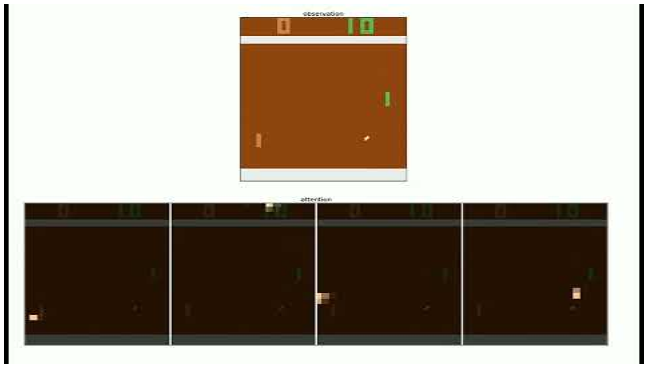
*请选择所有适用项。*

- [ ] Player
- [ ] Ball
- [ ] Opponent
- [ ] Score Board
- [ ] Some random, meaningless regions

**Explanation-Type 3: Attention Augmented (AA)**

Imagine the RL agent sitting in front of a TV and playing the game. The TV screen is shown at the top of the video. AA automatically identifies four visual concepts (which may not always be meaningful to humans) to focus on, as shown in the bottom row of the figure. **Your task is to evaluate each automatically identified concept from your perspective:** Does the automatically identified concept match any of the pre-defined concepts described in the Description Section (i.e., Player, Opponent, Ball, and Score Board)? If so, which specific pre-defined concept does it correspond to?

**YOU SHOULD:** Read through the questions and choices before watching the video.



4。 **[Single Selection]** According to the **1st** (start from left) explanation at the bottom row, what pre-defined concept do you think the RL agent is focusing on? **(Select "Hard to tell" if you cannot confirm decision in 5 seconds!)** *

*请仅选择一个答案。*

- ◯ Hard to tell
- ◯ Player
- ◯ Ball
- ◯ Opponent
- ◯ Score Board

5。 **[Single Selection]** According to the **2nd** (start from left) explanation at the bottom row, what pre-defined concept do you think the RL agent is focusing on? **(Select "Hard to tell" if you cannot confirm decision in 5 seconds!)** *

*请仅选择一个答案。*

- [ ] Hard to tell
- [ ] Player
- [ ] Ball
- [ ] Opponent
- [ ] Score Board

6。 **[Single Selection]** According to the **3rd** (start from left) explanation at the bottom row, what pre-defined concept do you think the RL agent is focusing on? **(Select "Hard to tell" if you cannot confirm decision in 5 seconds!)** *

*请仅选择一个答案。*

- [ ] Hard to tell
- [ ] Player
- [ ] Ball
- [ ] Opponent
- [ ] Score Board

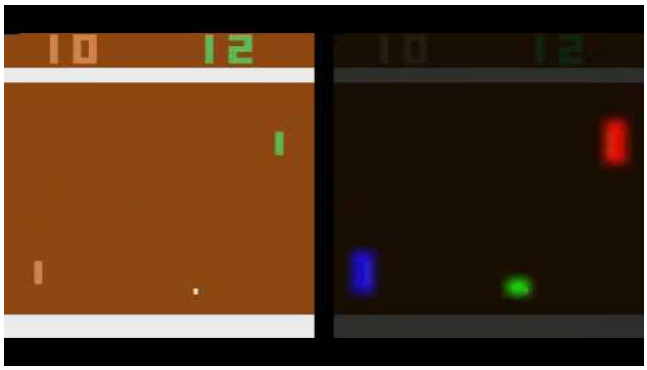
7。 **[Single Selection]** According to the **4th** (start from left) explanation at the bottom row, what pre-defined concept do you think the RL agent is focusing on? **(Select "Hard to tell" if you cannot confirm decision in 5 seconds!)** *

*请仅选择一个答案。*

- ◯ Hard to tell
- ◯ Player
- ◯ Ball
- ◯ Opponent
- ◯ Score Board

**Explanation-Type 4: Spatial Concept Transformer (SCT)**

**NOTE:** SCT for Pong is trained to focus on and correctly understand a subset of the pre-defined visual concepts: ***Player*** (Red), ***Ball*** (Green) and ***Opponent*** (Blue).



8。 **[Multiple Selection]** Based on the right side of the video, select one option for each bright region. *

*请选择所有适用项。*

- [ ] Player
- [ ] Ball
- [ ] Opponent
- [ ] Score Board
- [ ] Some random, meaningless regions

9。 **[Multiple Selection]** Among the pre-defined set of visual concepts that we wish the RL agent to foucs on, what concept is correctly understood?

*请选择所有适用项。*

- [ ] Select this option if the Player is marked in Red
- [ ] Select this option if the Ball is marked in Green
- [ ] Select this option if the Opponent is marked in Blue

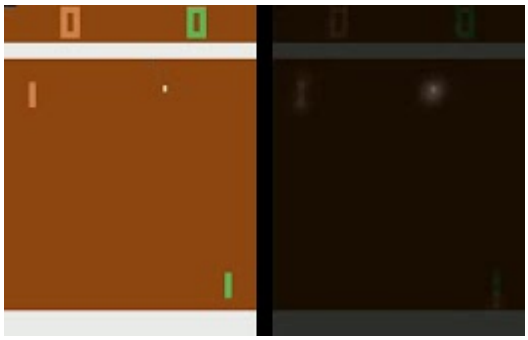
10。 **[Single Selection]** Which of the 4 types of explanation do you think best satisfies the following criteria:

1) It focuses on reasonable regions that align with your opinion.

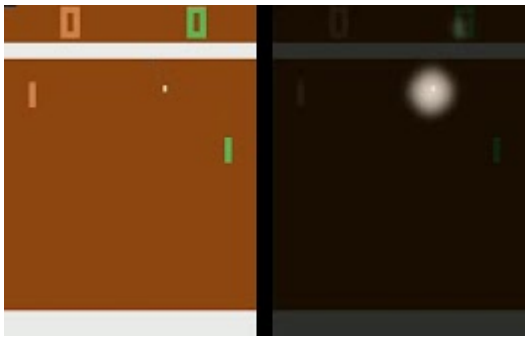
2) It effectively relates the focused regions to pre-defined concepts (i.e., Player, Opponent, Ball, and Score Board).

**NOTICE:** JSM and PSM do not satisfy criterion 2 because they do not relate the focused regions to any concepts, either automatically identified or pre-defined.

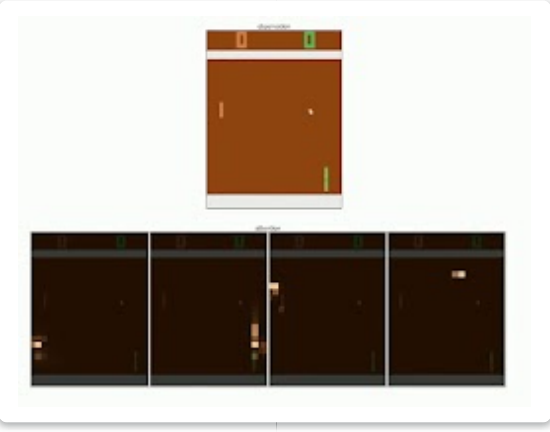
*请仅选择一个答案。*



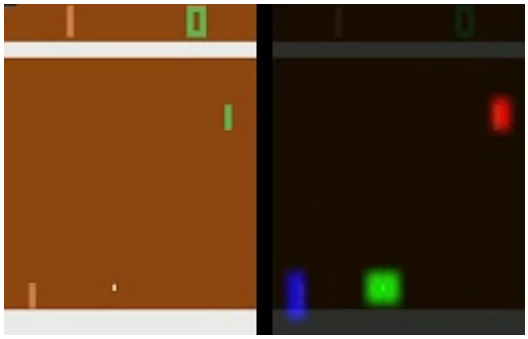
JSM



PSM



AA



SCT

## Game 2 Space-Invaders

Great job! You have finished questions for the game Pong. We move on to the 2nd game now.

**NOTE:** Some videos are long and you don't have to finish it. **Stop watching as long as you are ready for the questions.**

**!!!!!!!!!!!!!!!!!!!!!!  Some questions may have Multiple Selections !!!!!!!!!!!!!!!!!!!!!!**

**[Multiple Selection]:** multiple choice questions

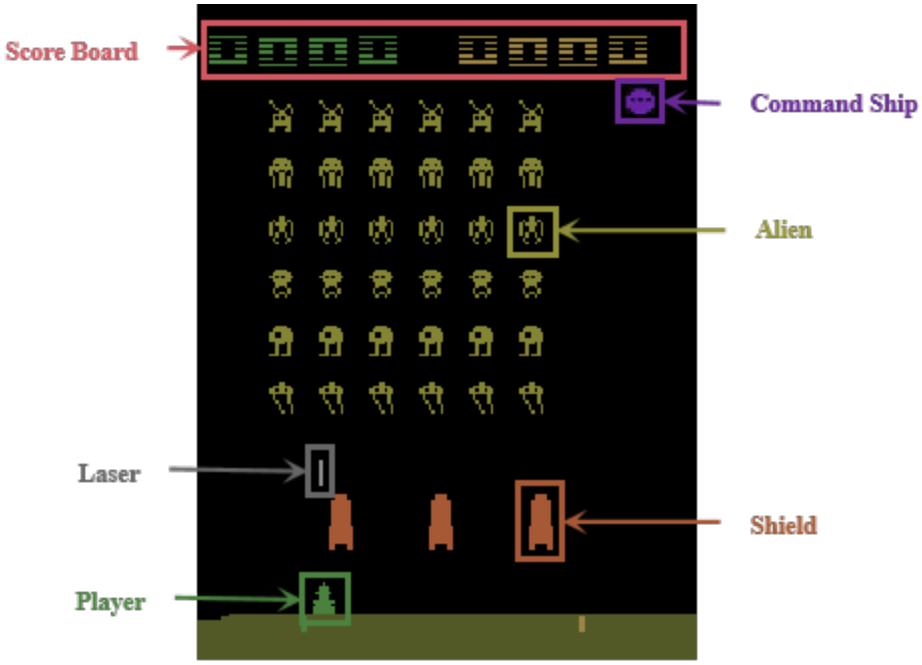
**[Single Selection]:** the question can have only one answer

### Description of the game

Each time you turn on SPACE INVADERS you will be at war with enemies from space who are threatening the earth. Your objective is to destroy these invaders by firing your "laser cannon.".

**Visual Concepts**

***Score Board:*** Displays the current score.
***Command Ship:*** A high-value target that gives significantly more points than any other target when hit.
***Alien:*** A regular target to destroy.
***Laser:*** Bullets emitted by the PLAYER.
***Shield:*** Initially, you are protected by the SHIELDS. However, as both you and the enemy hit the SHIELDS, they become damaged, allowing laser beams from your cannon and laser bombs from the enemy to pass through.
***Player:*** The spaceship controlled by either a human player or the RL agent.



11。 **[Multiple Selection]** What visual concept in this game do **YOU THINK** are most important while you are playing this game? ( A visual concept is considered crucial if removing it from the screen would hinder your ability to make effective **immediate reactions**.)

*请选择所有适用项。*

- [ ] Score Board
- [ ] Command Ship
- [ ] Alien
- [ ] Laser
- [ ] Shield
- [ ] Player

The RL agent may focus on areas that seem unreasonable to you. However, simply select the regions where the agent is objectively focusing. Please note, the following questions are **NOT** asking where you think the agent should focus, but where it is actually focusing.

**Explanation-Type 1: Jacobian-based Saliency Map (JSM)**

Imagine that the RL agent is sitting in the front of a TV. The left side of the video below shows the TV screen. The right side of the video below visualizes the RL agent's visual attention. Brighter regions indicate where the RL agent is focusing.

**YOU SHOULD:** Read through the questions and choices before watching the video.

12。 **[Multiple Selection]** Based on the right side of the video, select one option for each bright region.

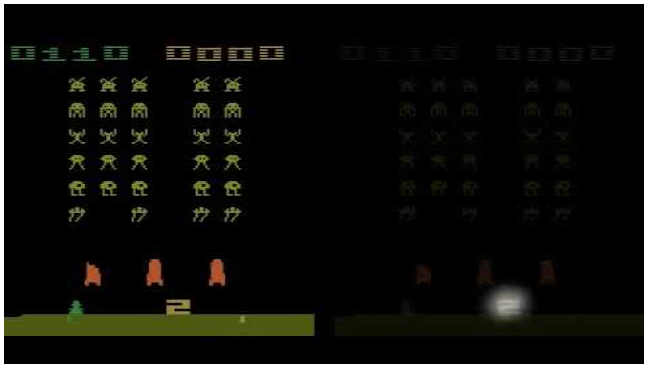
*请选择所有适用项。*

- [ ] Score Board
- [ ] Command Ship
- [ ] Alien
- [ ] Laser
- [ ] Shield
- [ ] Player
- [ ] Some random, meaningless regions

**Explanation-Type 2: Perturbation-based Saliency Map (PSM)**

Imagine that the RL agent is sitting in the front of a TV. The left side of the video below shows the TV screen. The right side of the video below visualizes the RL agent's visual attention. Brighter regions indicate where the RL agent is focusing.

**YOU SHOULD:** Read through the questions and choices before watching the video.



13。 **[Multiple Selection]** Based on the right side of the video, select one option for each bright region.

*请选择所有适用项。*

- [ ] Score Board
- [ ] Command Ship
- [ ] Alien
- [ ] Laser
- [ ] Shield
- [ ] Player
- [ ] Some random, meaningless regions

**Explanation-Type 3: Attention Augmented (AA)**

Imagine the RL agent sitting in front of a TV and playing the game. The TV screen is shown at the top of the video. AA automatically identifies four visual concepts (which may not always be meaningful to humans) to focus on, as shown in the bottom row of the figure. **Your task is to evaluate each automatically identified concept from your perspective:** Does the automatically identified concept match any of the pre-defined concepts described in the Description Section (i.e., Player, Shield, Laser, Alien, Command Ship and Score Board)? If so, which specific pre-defined concept does it correspond to?

**YOU SHOULD:** Read through the questions and choices before watching the video.

14。 **[Single Selection]** According to the **1st** (start from left) explanation at the bottom row, what pre-defined concept do you think the RL agent is focusing on? **(Select "Hard to tell" if you cannot confirm decision in 5 seconds!)**

*请仅选择一个答案。*

- ◯ Hard to tell
- ◯ Score Board
- ◯ Command Ship
- ◯ Alien
- ◯ Laser
- ◯ Shield
- ◯ Player

15。 **[Single Selection]** According to the **2nd** (start from left) explanation at the bottom row, what pre-defined concept do you think the RL agent is focusing on? **(Select "Hard to tell" if you cannot confirm decision in 5 seconds!)**

*请仅选择一个答案。*

- ◯ Hard to tell
- ◯ Score Board
- ◯ Command Ship
- ◯ Alien
- ◯ Laser
- ◯ Shield
- ◯ Player

16。 **[Single Selection]** According to the **3rd** (start from left) explanation at the bottom row, what pre-defined concept do you think the RL agent is focusing on? **(Select "Hard to tell" if you cannot confirm decision in 5 seconds!)**

*请仅选择一个答案。*

- ◯ Hard to tell
- ◯ Score Board
- ◯ Command Ship
- ◯ Alien
- ◯ Laser
- ◯ Shield
- ◯ Player

17。 **[Single Selection]** According to the **4th** (start from left) explanation at the bottom row, what pre-defined concept do you think the RL agent is focusing on? **(Select "Hard to tell" if you cannot confirm decision in 5 seconds!)**

*请仅选择一个答案。*

- ◯ Hard to tell
- ◯ Score Board
- ◯ Command Ship
- ◯ Alien
- ◯ Laser
- ◯ Shield
- ◯ Player

**Explanation-Type 4: Spatial Concept Transformer (SCT)**

**NOTE:** SCT is trained to focus on and correctly understand a subset of the pre-defined visual concepts: ***Player*** (Red), ***Shield***(Blue), ***Laser***(Yellow), ***Alien***(Green) and ***Command Ship***(Pink).

18。 **[Multiple Selection]** Based on the right side of the video, select one option for each bright region.

*请选择所有适用项。*

- [ ] Score Board
- [ ] Command Ship
- [ ] Alien
- [ ] Laser
- [ ] Shield
- [ ] Player
- [ ] Some random, meaningless regions

19。 **[Multiple Selection]** Among the pre-defined set of visual concepts that we wish the RL agent to foucs on, what concept is correctly understood?

*请选择所有适用项。*

- [ ] Select this option if the Player is marked in Red
- [ ] Select this option if the Shield is marked in Blue
- [ ] Select this option if the Laser is marked in Yello
- [ ] Select this option if the Alien is marked in Green
- [ ] Select this option if the Command Ship is marked in Pink

20。 **[Single Selection]** Which of the 4 types of explanation do you think best satisfies the following criteria:

1) It focuses on reasonable regions that align with your opinion.

2) It effectively relates the focused regions to pre-defined concepts (i.e., Player, Shield, Laser, Alien, Command Ship and Score Board).

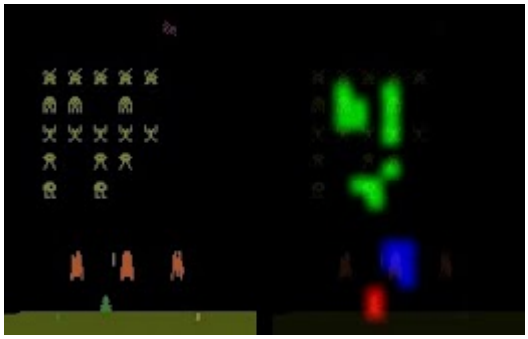
**NOTICE:** JSM and PSM do not satisfy criterion 2 because they do not relate the focused regions to any concepts, either automatically identified or pre-defined.

*请仅选择一个答案。*

JSM

PSM

AA



SCT

### Game 3 Battle-Zone

Greate! Here is the one last game to finish.

**NOTE:** Some videos are long and you don't have to finish it. **Stop watching as long as you are ready for the questions.**

**!!!!!!!!!!!!!!!!!!!!!!!!  Some questions may have Multiple Selections !!!!!!!!!!!!!!!!!!!!!!!!**

**[Multiple Selection]:** multiple choice questions

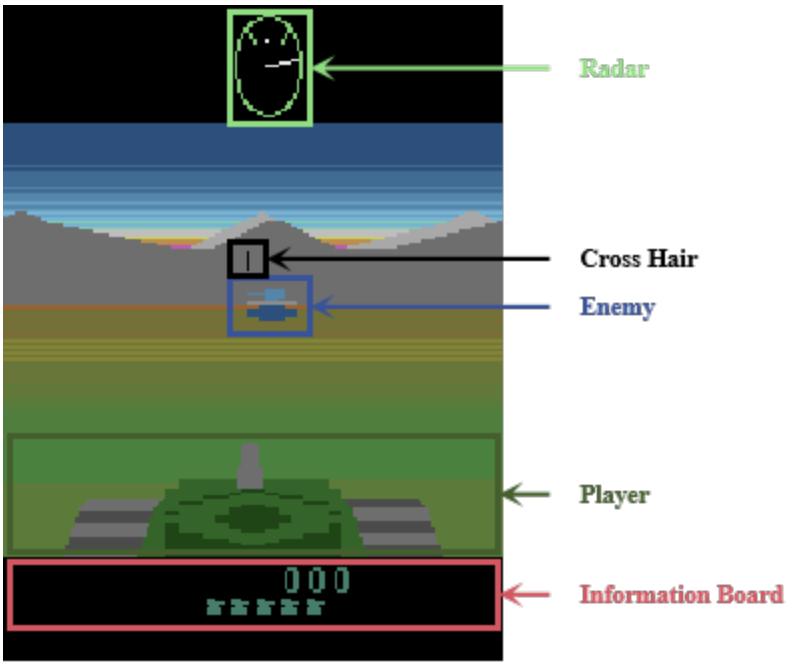
**[Single Selection]:** the question can have only one answer

### Description of the game

You control a tank and must destroy enemy vehicles. This game is played in a first-person perspective and creates a 3D illusion. A radar screen shows enemies around you.

**Visual Concepts**

***Radar:*** A screen displaying enemies (represented as white spots in the circle) around you. Radar is very useful for searching enimies when no enemy is currently in the front of the PLAYER.
***Cross Hair:*** A thin line that assists with aiming.
***Enemy:*** Enemy tanks that need to be destroyed.
***Player:*** The tank controlled by either a human player or the RL agent.
***Information Board:*** A panel showing the current score and remaining lives.




21。 **[Multiple Selection]** What visual concept in this game do **YOU THINK** are most important while you are playing this game? ( A visual concept is considered crucial if removing it from the screen would hinder your ability to make effective **immediate reactions**.)

*请选择所有适用项。*

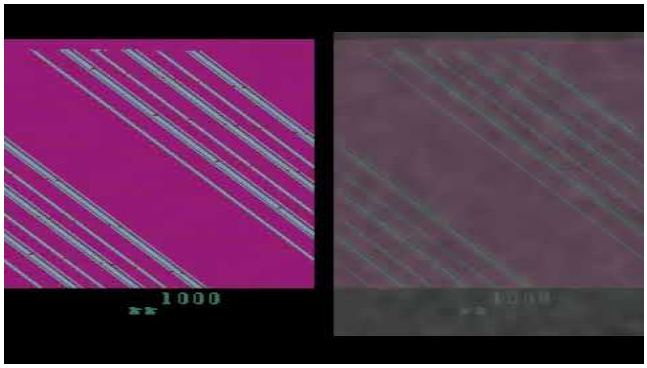
- [ ] Radar
- [ ] Cross Hair
- [ ] Enemy
- [ ] Player
- [ ] Information Board

The RL agent may focus on areas that seem unreasonable to you. However, simply select the regions where the agent is objectively focusing. Please note, the following questions are **NOT** asking where you think the agent should focus, but where it is actually focusing.

**Explanation-Type 1: Jacobian-based Saliency Map (JSM)**

Imagine that the RL agent is sitting in the front of a TV. The left side of the video below shows the TV screen. The right side of the video below visualizes the RL agent's visual attention. Brighter regions indicate where the RL agent is focusing.

**YOU SHOULD:** Read through the questions and choices before watching the video.




22。 **[Multiple Selection]** Based on the right side of the video, select one option for each bright region.

*请选择所有适用项。*

- [ ] Radar
- [ ] Cross Hair
- [ ] Enemy
- [ ] Player
- [ ] Information Board
- [ ] Some random, meaningless regions

**Explanation-Type 2: Perturbation-based Saliency Map (PSM)**

Imagine that the RL agent is sitting in the front of a TV. The left side of the video below shows the TV screen. The right side of the video below visualizes the RL agent's visual attention. Brighter regions indicate where the RL agent is focusing.

**YOU SHOULD:** Read through the questions and choices before watching the video.


23。 **[Multiple Selection]** Based on the right side of the video, select one option for each bright region.

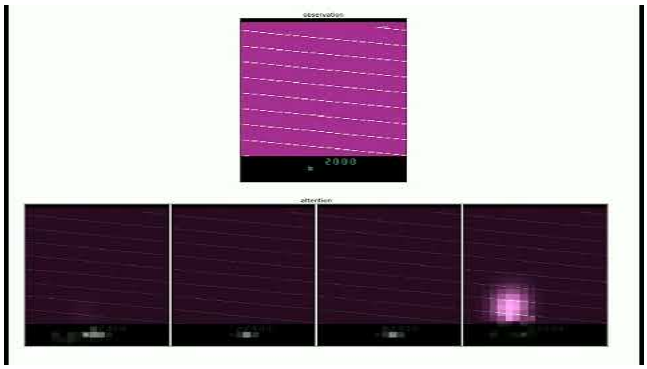
*请选择所有适用项。*

- [ ] Radar
- [ ] Cross Hair
- [ ] Enemy
- [ ] Player
- [ ] Information Board
- [ ] Some random, meaningless regions

**Explanation-Type 3: Attention Augmented (AA)**

Imagine the RL agent sitting in front of a TV and playing the game. The TV screen is shown at the top of the video. AA automatically identifies four visual concepts (which may not always be meaningful to humans) to focus on, as shown in the bottom row of the figure. **Your task is to evaluate each automatically identified concept from your perspective:** Does the automatically identified concept match any of the pre-defined concepts described in the Description Section (i.e., Radar, Cross Hair, Enemy, Player and Information Board)? If so, which specific pre-defined concept does it correspond to?

**YOU SHOULD:** Read through the questions and choices before watching the video.



24。 **[Single Selection]** According to the **1st** (start from left) explanation at the bottom row, what pre-defined concept do you think the RL agent is focusing on? **(Select "Hard to tell" if you cannot confirm decision in 5 seconds!)**

*请仅选择一个答案。*

- ◯ Hard to tell
- ◯ Radar
- ◯ Cross Hair
- ◯ Enemy
- ◯ Player
- ◯ Information Board

25。 **[Single Selection]** According to the **2nd** (start from left) explanation at the bottom row, what pre-defined concept do you think the RL agent is focusing on? **(Select "Hard to tell" if you cannot confirm decision in 5 seconds!)**

*请仅选择一个答案。*

- Hard to tell
- Radar
- Cross Hair
- Enemy
- Player
- Information Board

26。 **[Single Selection]** According to the **3rd** (start from left) explanation at the bottom row, what pre-defined concept do you think the RL agent is focusing on? **(Select "Hard to tell" if you cannot confirm decision in 5 seconds!)**

*请仅选择一个答案。*

- Hard to tell
- Radar
- Cross Hair
- Enemy
- Player
- Information Board

27。 **[Single Selection]** According to the **4th** (start from left) explanation at the bottom row, what pre-defined concept do you think the RL agent is focusing on? **(Select "Hard to tell" if you cannot confirm decision in 5 seconds!)**

*请仅选择一个答案。*

- ◯ Hard to tell
- ◯ Radar
- ◯ Cross Hair
- ◯ Enemy
- ◯ Player
- ◯ Information Board

**Explanation-Type 4: Spatial Concept Transformer (SCT)**

**NOTE:** SCT is trained to focus on and correctly understand a subset of the pre-defined visual concepts: ***Enemy*** (Red), ***Radar***(Green)**.**

28。 **[Multiple Selection]** Based on the right side of the video, select one option for each bright region.

*请选择所有适用项。*

- [ ] Radar
- [ ] Cross Hair
- [ ] Enemy
- [ ] Player
- [ ] Information Board
- [ ] Some random, meaningless regions

29。 **[Multiple Selection]** Among the pre-defined set of visual concepts that we wish the RL agent to foucs on, what concept is correctly understood?

*请选择所有适用项。*

- [ ] Select this option if the Enemy is marked in Red
- [ ] Select this option if the Radar is marked in Green

30。 **[Single Selection]** Which of the 4 types of explanation do you think best satisfies the following criteria:
1) It focuses on reasonable regions that align with your opinion.
2) It effectively relates the focused regions to pre-defined concepts (i.e., Radar, Cross Hair, Enemy, Player and Information Board) .

**NOTICE:** JSM and PSM do not satisfy criterion 2 because they do not relate the focused regions to any concepts, either automatically identified or pre-defined.

*请仅选择一个答案。*

JSM

PSM

AA

SCT