# Explaining Imitation Learning Through Frames

**Boyuan Zheng**[1], **Jianlong Zhou**[1], **Chunjie Liu**[2], **Yiqiao Li**[1], **Fang Chen**[1]

[1]University of Technology Sydney, Sydney
[2]Australian National University, Canberra

## Abstract

As one of the prevalent methods to achieve automation systems, Imitation Learning (IL) presents a promising performance in a wide range of domains. However, despite the considerable improvement in policy performance, the corresponding research on the explainability of IL models is still limited. Inspired by the recent approaches in explainable artificial intelligence, we proposed a model-agnostic explaining framework for IL models called R2RISE. R2RISE aims to explain the importance of frames with respect to the overall policy performance. It iteratively retrains the black-box IL model from the randomized masked demonstrations and uses the conventional evaluation outcome environment returns as the coefficient to build an importance map. We also conducted experiments to investigate three major questions concerning frames' importance equality, the effectiveness of the importance map, and connections in importance maps from different IL models. The result shows that R2RISE distinguishes important frames from the demonstrations effectively. Code is available from https://anonymous.4open.science/r/ExIL.

## Introduction

Recent advances in Imitation Learning (IL), which leverages external demonstrations to reproduce the desired behaviours, demonstrate a promising performance in fields like 3D gameplay (Scheller, Schraner, and Vogel 2020), robotics (Yu et al. 2018), and automatic driving (Codevilla et al. 2019). Despite their success, most research in IL focuses on applying complex Deep Neural Network (DNN) models, such as convolutional neural networks (CNN) and generative adversarial networks (GAN), to achieve high performance across different conditions. However, less attention is given to explaining what information the trained agents have learned from the external demonstrations. Consequently, IL methods are increasingly becoming less interpretable, posing an open challenge in combining IL and Explainable Artificial Intelligence (XAI).

On the other hand, XAI has garnered attention from the research community in recent years, particularly in the field of computer vision. Methods like LIME have been proposed to explain image predictions (Ribeiro, Singh, and Guestrin 2016). As the concept of explainability gradually spreads to other domains, it has also found its way into reinforcement learning. For instance, Shu et al. (2017) introduced hierarchical policies to explain complex tasks by decomposing top-level policies into several lower-level actions. Madumal et al. (2020) aim to explain model-free reinforcement learning using causality models to address questions like "Why (not) action A?" Xie et al. (2022) introduce an innovative framework that employs one of IL methods called Adversarial Inverse Reinforcement Learning to furnish comprehensive explanations for the decisions made by a reinforcement learning model and captures the model's intuitive tendencies by summarizing its decision-making process. In contrast, research investigating the combination of explainability and imitation learning is relatively recent.

Pan et al. (2020) proposed a model-specific method called xGAIL, which leverages existing XAI methods to explain single action predictions made by the state-of-the-art IL method, Generative Adversarial Imitation Learning (GAIL) (Ho and Ermon 2016). Before xGAIL's introduction, the IL research community had explored features in image inputs, but the significance of explainability had not been explicitly highlighted. For example, Brown et al. (2019) utilized attention maps of input image frames to validate the effectiveness of the learning process. De Haan et al. (2019a) pointed out that IL agents could learn incorrect causal correlations between expert behaviors and irrelevant input features. These methods, including xGAIL, demonstrate the significance of features in individual image frames within trajectories for models to learn desired behaviors. However, most IL methods are evaluated based on policy performance in the environment rather than analysing a single behavior prediction for a given state, and the aforementioned methods do not assess the importance of individual frames within input trajectories. The question arises: do the input image frames have identical importance, and if not, how can frames' importance be distinguished concerning overall policy performance?"

To tackle these problems, we attempt to explain the input demonstrations as a whole by proposing a novel explaining method called Remove and Retrain via Randomized Input Sampling for Explanation (R2RISE), which iteratively masks random frames in the demonstrations and evaluates the performance of the agents trained by the masked inputs. The intuition is that the input demonstrations are regarded as a single image, and frames in the demonstrations are regarded as pixels. In this case, existing XAI and computer vision methods could be directly applied to investigate the importance of frames instead of features in a single frame. R2RISE combines the existing methods RISE (Petsiuk, Das, and Saenko 2018) and ROAR (Hooker et al. 2019), and achieves model-agnostic explanations for IL models with various architectures.

Our main contribution is summarized as follows: 1) We propose a model-agnostic method to explain IL models; 2) We extend a novel perspective to explain IL with respect to the whole input dataset instead of a specific frame; 3) We investigate the connection between agents' overall performance and demonstration frames;

## Preliminaries

To better illustrate our approach, we first introduce the related literature in XAI: RISE (Petsiuk, Das, and Saenko 2018) and ROAR (Hooker et al. 2019), followed by a review of existing research related to explainable imitation learning.

### Randomized Input Sampling for Explanation (RISE) & RemOve And Retrain (ROAR)

Randomized Input Sampling for Explanation (RISE) is a state-of-the-art XAI method proposed by Petsiuk et al. (2018) to explain black-box models in image classification. RISE stands out for its simplicity and generality. Unlike other popular XAI approaches that rely on gradient calculations of image classification outputs, RISE probes the target model by randomly masking the input image under a predefined degradation level and recording the resulting probability for the target class. This process is repeated multiple times, and the recorded probabilities for each pixel are linearly combined to generate an importance map. This map identifies the most influential regions in the input image for the target decision. In the context of IL, which typically requires multiple demonstrations to train the model, we can consider whole demonstrations as a single image where the frames could be regarded as pixels. Here, the trajectory's length becomes the image width, and the number of trajectories represents the image height. In this regard, RISE becomes significant for explaining IL and identifying the most influential frames for policy training. However, using the output probability as the coefficient to accumulate the importance map does not align with the traditional IL evaluation approach, where the overall environment return is used to assess the model's performance. To adapt RISE for conventional IL evaluation, optimizations are required.

On the other hand, RemOve And Retrain (ROAR), proposed by Hooker et al., offers a reliable evaluation of feature importance for a wide range of XAI methods. By substituting certain pixels, estimated to be important, with fixed uninformative values and then retraining a new model, ROAR determines the model's sensitivity to pixel removal. If the model demonstrates a sharp degradation in performance due to the removal, it suggests that the proposed model is more accurate. The authors argue that the retraining process is essential to ensure low variance in performance, as machine learning models commonly assume similarity between training and test distributions. However, it is important to note that ROAR serves as an evaluation framework for XAI methods, aiming to achieve a more robust evaluation of these methods. Nevertheless, ROAR itself does not possess the capability to determine feature importance directly. Consequently, applying ROAR directly to explain IL becomes unfeasible. However, we can draw on the intuition of ROAR, which involves retraining several models under the same removal rate, to obtain more robust results in the context of IL explanation. Moreover, since conventional IL evaluation involves representing the performance of the trained model as returns from a dynamic environment, training the model once with fixed image observations and masks is not appropriate. The ROAR framework allows for masking the samples before training the model, enabling the explanation of IL models while utilizing conventional evaluation methods.

### Explainable Imitation Learning

The literature on the combination of XAI and IL remains limited at present (Zheng et al. 2022). They can be broadly categorized into two approaches aimed at achieving better explainability. The first approach involves leveraging white-box models, wherein existing neural network architectures are replaced with models possessing intrinsic interpretability. Alternatively, the learned policy is represented in a hierarchical structure to enhance interpretability. For instance, Leech (2019) proposed a learning framework that combines IL with logical automata, representing problems as compact finite state automata with human-interpretable logic states. Bewley et al. (2020), on the other hand, modeled the behavior policy of a trained black-box agent using a decision tree generated from analyzing its input-output statistics. Zhang et al. (2021) leveraged a hierarchical framework to decompose the complex task, explaining the model's decision-making process and analyzing the causes of failure.

On the other hand, research related to analyzing pixel-wise explainability focuses on CNN structures in IL models, which are widely used to capture features from image inputs. Drawing from existing studies in XAI and computer vision, model explainability is commonly represented as heatmaps, enabling the analysis of the model's decision-making process. For instance, Pan et al. (2020) endeavored to explain the state-of-the-art GAIL model (Ho and Ermon 2016) through a model-specific explanation method called xGAIL. Their approach was validated on a passenger-seeking problem that involved spatial-temporal data, successfully elucidating individual decision-making based on extracted frames. While xGAIL provided local and global explanations for a well-trained GAIL model, it possesses certain limitations. Firstly, Its input data type is restricted to geographical data, and the explanation framework is less likely to be applicable to other IL models since xGAIL is model-specific. Furthermore, xGAIL transforms the IL problem into an image classification problem, resulting in the extraction and analysis of limited frames from abundant inputs. Consequently, individual decision-making is evaluated from a single image (frame), which might not offer an intuitive understanding of the IL problem in its entirety. In a dynamic environment, this approach might lack a comprehensive overall explanation with respect to the policy performance, and the generated explanations may be biased as

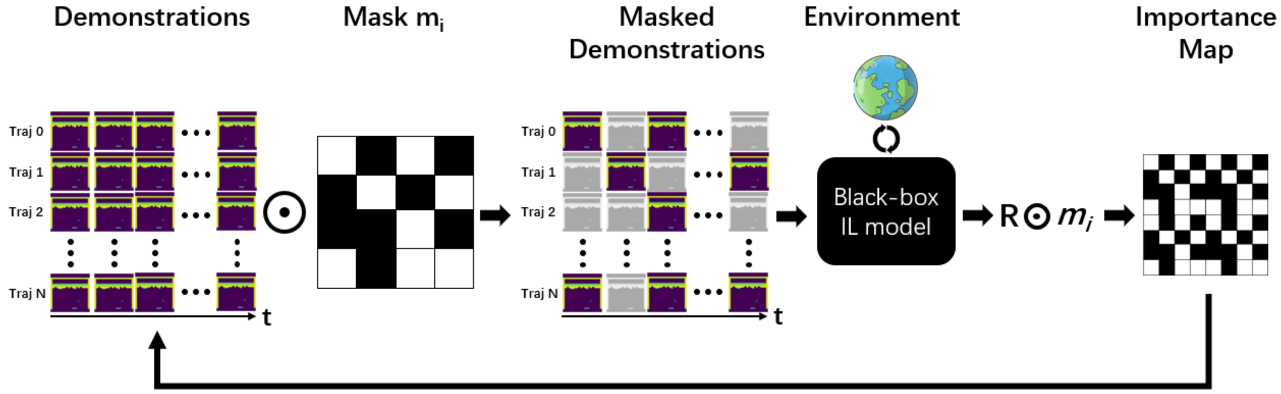


Figure 1: A diagrammatic representation of a single iteration of R2RISE. The input demonstrations are subject to element-wise multiplication (denotes as $\bigodot$) with a random mask which creates a masked demonstration, with greyed frames indicating those which are masked. Subsequently, the masked demonstration is used to train a black box IL model. The trained model interacts with the test environment to obtain returns, the mean of which is element-wise multiplied with the initial mask and accumulated to the existing importance map.

a substantial amount of information from the demonstrations gets filtered out during frame extraction.

## R2RISE

To overcome the above-mentioned limitations, we propose a model-agnostic explanation method for imitation learning called R2RISE. In the interest of maintaining generality, we do not introduce the Markov Decision Process in this paper, as it is unnecessary for those model-free IL methods that do not evaluate the environment dynamics. R2RISE combines the merits of RISE and ROAR, aligning with prevalent IL problem settings and examining the importance of frames in relation to overall policy performance.

We first review how to distinguish pixels' importance using randomized masks for the image classification problem. For a given image $\mathcal{I}$ with the size of $H \times W$, we create a random binary mask $m$ with the same size of $\mathcal{I}$ and do an element-wise multiplication between image $\mathcal{I}$ and mask $m$ (denoted as $\mathcal{I} \bigodot m$). The masked images are then fed into the black-box model (denoted as $f(\mathcal{I} \bigodot m)$). The importance of pixels is defined as the expected score over all possible masks $M = \{m_0, m_1, ..., m_i\}$ conditioned on the event that pixel is observed (denoted as $M(\lambda) = 1$, if the pixel is masked, then $M(\lambda) = 0$), i.e. $S_{\mathcal{I},f}(\lambda) = \mathbb{E}_M[f(\mathcal{I} \bigodot m)|M(\lambda) = 1]$. By rewriting the above equation as a summation over mask $m$ and empirically estimating it using Monte Carlo sampling, the saliency map can be computed as a weighted sum of random masks and normalized by the expectation of $M$:

$$S_{\mathcal{I},f}(\lambda) \approx \frac{1}{\mathbb{E}[M] \cdot N} \sum_{i=1}^{N} f(\mathcal{I} \bigodot m_i) \cdot m_i(\lambda). \quad (1)$$

Since the above formulation does not need any assumptions or information from the target model, this could be used to explain black-box models. The intuition is that when $f(\mathcal{I} \bigodot m)$ is high, it indicates that the mask observes important pixels. However, directly applying the above method to IL is inappropriate since applying a fixed mask on a dynamic environment frame-by-frame is not reasonable. Additionally, IL methods commonly evaluate policy networks through interactions with the environment rather than feeding them with another dataset. In this case, if the model needs to be well-trained in advance, the conventional IL evaluation method becomes inapplicable, making it impractical to obtain frame-wise importance for all input trajectories. To address the aforementioned issues, we draw inspiration from the concept of ROAR, which can not distinguish the feature importance but allows masking samples before training the model. Our approach involves retraining multiple models using diverse masked datasets, and we accumulate the environment returns of these models to create a frame-wise importance map.

Like most imitation learning methods, we assume the testing data has a similar distribution as training data, and the input demonstrations $\mathcal{D}_n$ are optimal. This could ensure evaluation fairness for a wide range of IL models. The demonstrations $\mathcal{D}_n$ consist of multiple trajectories, and each trajectory could be represented as either a sequence of state-action pairs or observations. In this work, we represent the trajectory as a sequence of state-action pairs, i.e. $\mathcal{D}_n = \{\tau_1, \tau_2, ..., \tau_n\}$, where $\tau_{i \in [1,n]} = \{(s_1, a_1), (s_2, a_2), ..., (s_t, a_t)\}$. The black-box imitation learning model trains a policy (denoted as $\pi_{\mathcal{D}_n}(a|s)$) on the input demonstrations $\mathcal{D}_n$, then interacts with the environment and obtains returns $R$. For the finite horizon T, the expected return could be represented as the accumulation of the return at each time step, i.e.

$$R(\pi_{\mathcal{D}_n}) = \mathbb{E}[\sum_{t=0}^{T} r_t | \pi_{\mathcal{D}_n}]. \quad (2)$$

In this work, we assess the model's performance by measuring its cumulative environment returns instead of the faithfulness of the learned policies compared with the demonstrators, which aligns with the prevalent evaluation approaches used in most IL methods and ensures the generality of R2RISE.

The discussion we have so far motivated us to propose a frame-wise explanation method for IL called R2RISE. By regarding the demonstrations $\mathcal{D}$ as a single image, where the number of demonstrations is the image height $\mathcal{H}$, and the length of the demonstration is the image width $\mathcal{T}$, we could investigate the frame-wise importance by iteratively applying numerous randomized masks on the demonstrations and accumulating the environment returns from the black-box IL models trained on the masked demonstration. R2RISE hypothesises that the importance of each frame is not identical and iteratively removes random frames based on the predefined degradation level. The modified dataset $D_n = D \bigodot m_i$ is used to retrain an IL model. The retrained IL model then constantly interacts with the environment to obtain the accumulative return, and R2RISE finally compute the linear combination of the returns to obtain the saliency map (See Figure 1). Assuming the number of generated masks is $N$, and the return of each mask is the average return from $J$ rounds of interaction with the environment, the computation of the saliency map is similar to equation (1). We also partitioned the trajectories into snippets to decrease the number of frame combinations. This enabled us to repeat distinct combinations multiple times under the high degradation level, thus amplifying the importance between frames. To cater to the setting of IL, we substitute the $f(\mathcal{I} \bigodot m)$ in equation (1) with equation (2):

$$S_{\mathcal{D}_n,f}(\lambda) \approx \frac{1}{\mathbb{E}[M] \cdot N} \sum_{i=1}^{N} R(\pi_{\mathcal{D}_i}) \cdot m_i(\lambda) \tag{3}$$

$$= \frac{1}{\mathbb{E}[M] \cdot N} \sum_{i=1}^{N} \mathbb{E}[\sum_{t=0}^{T} r_t | \pi_{\mathcal{D}_i}] \cdot m_i(\lambda) \tag{4}$$

$$= \frac{1}{\mathbb{E}[M] \cdot N \cdot J} \sum_{i=1}^{N} \sum_{j=0}^{J} \sum_{t=0}^{T} r_t \cdot m_i(\lambda) \tag{5}$$

where $\mathcal{D}_i = \mathcal{D} \bigodot m_i$, and

$$m_i(\lambda) = \begin{cases} 0, & \text{if the frame is masked,} \\ 1, & \text{if the frame is observed.} \end{cases}$$

As the formula presented, R2RISE also does not require any information from the IL models, such that R2RISE could be used as a model-agnostic method to explain IL. We summarize the above process of R2RISE in Algorithm 1.

To evaluate the effectiveness of R2RISE, we conduct tests using three diverse IL methods: Behavioral Cloning (BC) (Bain and Sammut 1999), Generative Adversarial Imitation Learning (GAIL) (Ho and Ermon 2016), and Behavioral Cloning from Observation (BCO) (Torabi, Warnell, and Stone 2018). In BC, the control policy is obtained under a supervised learning fashion, directly mapping states to actions. On the other hand, GAIL learns the policy through an iterative adversarial process involving a generator G that produces fake data distributions and a discriminator D that distinguishes between the fake data distribution and the expert distribution. In contrast to the previous two methods, BCO does not require action labels. It employs an inverse dynamic model to estimate actions from two adjacent input image frames and iteratively optimizes both the policy network and the inverse dynamic model. These methods differ significantly in how they learn the policy, the network structures they employ, and BCO even uses a different type of input. Our objective is to validate the generality of R2RISE across this diverse model selection.

Algorithm 1: R2RISE

**Input**: demonstration dataset $\mathcal{D}$, target IL model $f$
**Parameter**: number of randomized masks $N$, degradation level $l$, size of each grid in mask $z$
**Output**: an importance map $S_{\mathcal{D},f}$

```
Initialize masks $M$ based on the number of randomized masks $N$, degradation level $l$ and size of each grid in mask $z$.
Initialize blank importance map $S_{\mathcal{D},f}$ with the same shape as $D$.
for $m_i$ in M do
    Randomly initializes the black-box model $f$.
    Obtain masked demonstrations by element-wise multiplication $D_n = D \bigodot m_i$.
    Train black-box model $f$ with the masked demonstrations $D_n$ and obtain policy $\pi_{D_n}$.
    Evaluate policy $\pi_{D_n}$ by interacting with environment repeatedly and obtain average return $\bar{R}$.
    Update importance map via element-wise addition, $S_{\mathcal{D},f} \leftarrow S_{\mathcal{D}_n,f} \bigoplus (\bar{R} \bigodot m_i)$
end for
return importance map $S_{\mathcal{D},f}$
```

## Experiment

In this section, we conduct a series of experiments and address the following questions: (1) Is the importance between frames identical? (2) Can R2RISE distinguish the importance between frames? (3) Are there connections between the importance map obtained from different IL models?

### Setup

We implemented experiments with NVIDIA Quadro RTX 5000 GPU, and three different IL models, BC, GAIL, and BCO, were evaluated on three OpenAI Gym Atari tasks: BeamRider, Breakout, and SpaceInvaders(Brockman et al. 2016). In BeamRider and SpaceInvaders, the agent controls a warplane to shoot down enemies while avoiding their attacks; in Breakout, the agent operates a bottom paddle to bounce a ball and break the upper bricks.

Similar to recent IL methods, we leverage the proximal policy optimization (PPO) (Schulman et al. 2017) algorithm from the OpenAI baselines (Dhariwal et al. 2017), utilizing default parameters and reward function, to generate expert demonstrations. The observations with the size of $84 \times 84 \times 3$ and actions between the PPO agents and the task environment are recorded as “trajectories.” These trajectories, generated from checkpoint 1400, serve as expert demonstrations. To avoid the “causal confusion” problem (models build wrong causal relationships with irrelevant patterns) (De Haan, Jayaraman, and Levine 2019b) and ensure the fairness of our evaluation, we mask the indicators (such as scoring broad) in frames and ensure the same demonstrations as input for different IL models.

Regarding the parameter setting, we generate 20 trajectories with a fixed length of 1000 for each IL model. Five random seeds and five levels of percentage degradations $l = [10, 30, 50, 70, 90]$ are pre-defined for evaluation. We propose to retrain 100 models with 100 different randomized masks. Each mask contains 20*100 grids, which means that every single trajectory is cut into 100 snippets, and each snippet assigns the same importance to 10 continuous frames. The retrained model is tested from 20 trials, and the average return of each trial, multiplied element-wisely with the random mask, is then linearly combined to generate the accumulated importance map.

## Is the importance between frames identical?

Remember that we hypothesize that the importance of frames is different, so we validate this hypothesis by applying several randomized masks on the same demonstration and comparing the performance of the trained model. If the outcomes present noticeable deviations, then it can be inferred that the contributions between frames vary. To further this idea, we divide each trajectory into ten segments of equal length and randomly assign either a value of 0 or 1 to each segment. We control the number of 1s and 0s to ensure equal input data amount across trials. Regions assigned 0 are removed, and then the preprocessed demonstrations are used as input to train the relevant models. The trained model’s performance was evaluated based on the average environment returns from ten trials. This process iterates ten times, and we get Table 1. Here, the numbers outside the brackets represent the average environment returns from 20 evaluation episodes, while the numbers inside the brackets denote the standard deviation.

Table 1: Variation in Model Performance Across Ten Trials.

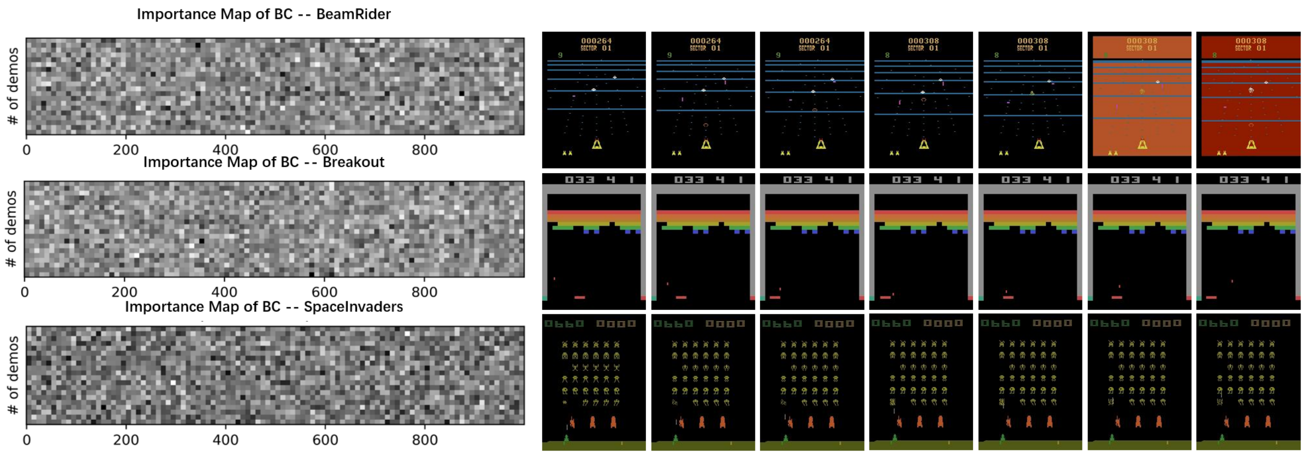
| | | Beamrider | Breakout | SpaceInvaders |
|---|---|---|---|---|
| expert demo | | 735.0 (115.29) | 34.7 (5.91) | 679.5 (102.29) |
| BC | min | 573.2 (216.21) | 4.2 (3.24) | 330.8 (168.46) |
| | medium | 641.0 (151.82) | 13.1 (5.95) | 424.5 (183.23) |
| | max | 732.1 (209.66) | 25.8 (10.01) | 529.5 (160.01) |
| | without M | 1884.2 (670.66) | 15.6 (6.83) | 486.7 (197.14) |
| GAIL | min | 290.4 (167.77) | 3.3 (2.60) | 275.6 (99.17) |
| | medium | 382.8 (126.78) | 4.6 (2.82) | 353.1 (160.61) |
| | max | 440.8 (176.73) | 6.8 (2.11) | 441.3 (201.72) |
| | without M | 387.5 (156.95) | 8.1 (2.75) | 361.6 (135.6) |
| BCO | min | 88.0 (59.03) | 2.9 (2.56) | 59.5 (34.02) |
| | medium | 333.3 (139.22) | 8.0 (4.49) | 117.5 (85.69) |
| | max | 710.4 (203.06) | 13.4 (5.69) | 460.4 (74.18) |
| | without M | 598.0 (136.57) | 16.8 (8.57) | 203.0 (59.42) |

[a] Without M denotes policy performance without masking.

From Table 1, it becomes evident that performance deviations exist not only across different models (ANOVA: Beamrider: F=14, $p < .000$; breakout: F=14.02, $p < .000$; SpaceInvaders: F=15.69, $p < .000$) but also from trial to trial. For a given model type, the performance spectrum spans significantly from the best to the worst, indicating that the frames do not have identical contributions toward the policy performance. Compared to GAIL and BC, the performances of BCO fluctuate significantly, which indicates BCO is more sensitive to the important frames than the other two algorithms. Notably, in the context of the BeamRider task, both BC and GAIL exhibit analogous trends when applying identical masks on input trajectories, prompting inquiries into potential correlations among importance maps derived from diverse models. Upon confirming inherent frame disparities, we introduce R2RISE to extract importance maps for given trajectories. Figure 2a depicts importance maps obtained through BC (with additional algorithms detailed in the supplemental material). The x-axis denotes trajectory length, while the y-axis represents trajectory count, and grayscale shading indicates relative importance. A lighter shade implies higher importance. Likewise, the importance maps generated by R2RISE highlight frame disparities, consistent with Table 1 findings. However, there is no discernible pattern, suggesting a discrete rather than clustered distribution of important frames along trajectories. Figure 2b illustrates the frames that have been identified as important components within the demonstrations. In the context of BeamRider and SpaceInvaders, models prioritize actions related to destroying enemy flights, while Breakout models emphasize rebounding of upper blocks, sidewalls, and the paddle. These extracted frames provide a lens through which the model’s learning process can be investigated, enhancing comprehension of the underlying “causes” influencing the overall policy performance.

## Can R2RISE distinguish the importance between frames?

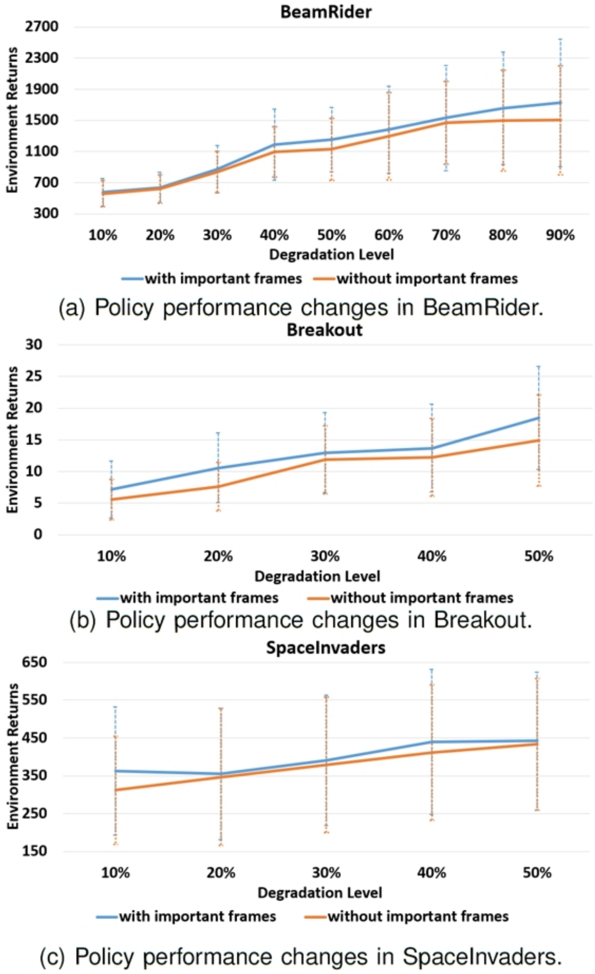
This section investigates the effectiveness of R2RISE. The abovementioned hypothesis validation indicates that the importance between frames varies. To obtain a map indicating the importance of frames, we implement R2RISE. However, the quality of the generated maps needs to be properly evaluated. In this case, we adopt similar causal metrics used by Petsiuk et al. (2018), insertion and deletion, where the availability of the ‘cause’ will significantly influence the model’s decision-making and performance. Under the scenario of image classification tasks, deleting the causal pixels will lead to a sharp drop in accuracy if the model gets well explained. In our experiment, we leverage similar intuition and expect the removal of the important frames would lead to a worse performance while limiting the amount of input data to be the same. To achieve this, we transform the generated importance map into a mask according to different degradation levels and replace the map with either 1s or 0s, depending on the degradation level.



(a) Importance maps of BC obtained by R2RISE.

(b) Frames identified as important in the Atari domain.

Figure 2: Importance maps and the corresponding extracted sample frames that are recognized as important.



(a) Policy performance changes in BeamRider.

(b) Policy performance changes in Breakout.

(c) Policy performance changes in SpaceInvaders.

Figure 3: Validations of the effectiveness of R2RISE. The lines and the error bars represent the mean performance and standard deviation of the model trained from a certain percentage of the most important (or least important) frames.

Figure 3 shows the changes in policy performance using different percentages of the most important (or least important) frames. The x-axis is the percentage of data used to train the model, and the y-axis is the environment returns. The solid lines are the average returns from 10 trials each with 20 evaluation episodes using the same transformed mask and demonstrations, the error bar is the standard deviation. From Figure 3, we can observe that the models trained with the most important frames perform better when the input data is relatively limited for all tasks (BeamRider (50%): t=2.9971, $p < .01$; Breakout (50%): t=4.6522, $p < .01$; SpaceInvaders (10%): t=3.2414, $p < .01$), which meets our expectations. In addition, with more training data fed into the network, the standard deviation enlarges, likely due to the presence of redundant information, the informativeness of frames varies even within the same grid. In the context of the BeamRider task (refer to Figure 3 (a)), the model with important frames performs better than the model trained with the least important frames. The performance deviation at the beginning of the figure is relatively small, we think the reason is that the model is more sensitive to the amount of data than to the availability of the important frames. From the end of this figure, it can be seen that the performance deviation increases, indicating that the identified top important frames significantly determine the upper bound of the model's performance and removing these top influential frames results in a sharp decrease. For task Breakout and SpaceInvaders (see Figure 3 (b) and 3 (c)), it can be seen that the model trained with important frames at least slightly outperforms its counterpart lacking such frames, particularly in scenarios with limited input data. This observation underscores the extraction of more valuable knowledge from frames identified as important, thus confirming the effectiveness of R2RISE. However, as the dataset size increases, the performance of the two models begins to converge. We suspect this phenomenon is attributed to the addition of more ordinary or redundant frames to the dataset, potentially leading to the convergence or even negative im-

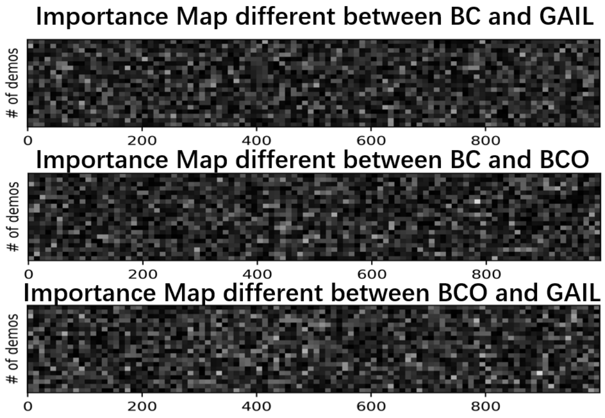


Figure 4: Deviation between importance maps in Breakout.

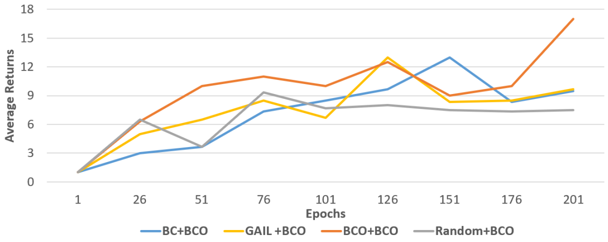


Figure 5: Average reward learning curves of GAIL trained with different masks. The blue, orange and grey lines are trained with the masks extracted from the importance map obtained using BC, GAIL and random, respectively.

pact on policy performance.

### Are there connections between the importance map obtained from different IL models?

Remember that we observed a similar trend between GAIL and BC in BeamRider tasks when applying identical masks on input trajectories, this prompt us to examine the connections between IL models. In this section, we investigate the question: does the importance map obtained by one model have connections to another model? To this end, we propose two approaches to explore intrinsic connections between models. The first attempt directly compares the importance maps by projecting the values into the same range and calculating element-wise deviation (see Figure 4). The larger the deviation is, the whiter the output image should be. From Figure 4, we can observe that most grids are close to black, which indicates both models assigned similar importance to these grids. The second attempt involves generating a mask from an importance map generated by one model and applying that mask to the same demonstrations to train another model. The underlying assumption is that if there are connections between models' importance maps, the important frames identified by one model should work well on another model, leading to improved performance compared to a randomized mask on the frames. Figure 5 displays the average returns of BCO using four types of masks, it can be seen that when applying the mask obtained from BCO, the model BCO performs better across most checkpoints compared to other masks, which suggests R2RISE does distinguish important frames from the input trajectories. Regarding masks derived via BC and GAIL, we can see the performance is still better than the randomized masks, this highlights a degree of similarity in the importance maps obtained from different IL models, which is consistent with the conclusion drawn from the first attempt.

## Limitations

Several intriguing challenges await further exploration. Currently, the framework relies on a single trial to train the model, achieving robustness through substantial randomized masks. Ensembling IL models could further ensure low variance for each given mask. In addition, computation intensiveness is another limitation. The performance and robustness of R2RISE heavily rely on the number of masks used to generate the importance map, so that the time taken to obtain a satisfactory explanation is closely linked to the time spent on training the target model once. If the target model requires days to train, it would not be practical to retrain hundreds of times. Improving time efficiency while preserving the model-agnostic property remains an open challenge for R2RISE. Investigating the relationship between global explanations and the frames that are recognized as important is another interesting future direction. Although we observed similar patterns in the extracted frames from different trials and models, it is still unsafe to claim these patterns could be the global explanations for a given task. Further research is needed to provide theoretical proof for the connections.

## Conclusion

This paper introduced a model-agnostic explaining framework for imitation learning called R2RISE. It distinguishes the frames' importance in relation to the overall policy performance. It iteratively applies numerous randomized masks on the demonstrations and retrains the black-box IL model from the masked demonstration. Similar to conventional IL methods, model evaluation is measured via accumulated returns from the environment. We leverage these accumulated returns as a coefficient to multiply with the mask and linearly combine the multiplied masks to obtain the importance map of the frames. Experiments have shown that the importance of frames is not equal, and R2RISE can successfully distinguish important frames from the demonstrations, offering valuable insights to probe IL models for better explanations.

## References

Bain, M.; and Sammut, C. 1999. A framework for behavioural cloning. In *Machine Intelligence 15*, 103–129. Oxford University Press.

Bewley, T.; Lawry, J.; and Richards, A. 2020. Modelling agent policies with interpretable imitation learning. In *International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning*, 180–186. Springer.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

Brown, D.; Goo, W.; Nagarajan, P.; and Niekum, S. 2019. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, 783–792. PMLR.

Codevilla, F.; Santana, E.; López, A. M.; and Gaidon, A. 2019. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9329–9338.

De Haan, P.; Jayaraman, D.; and Levine, S. 2019a. Causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 32.

De Haan, P.; Jayaraman, D.; and Levine, S. 2019b. Causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 32.

Dhariwal, P.; Hesse, C.; Klimov, O.; Nichol, A.; Plappert, M.; Radford, A.; Schulman, J.; Sidor, S.; Wu, Y.; and Zhokhov, P. 2017. Openai baselines.

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

Hooker, S.; Erhan, D.; Kindermans, P.-J.; and Kim, B. 2019. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32.

Leech, T. 2019. *Explainable machine learning for task planning in robotics*. Ph.D. thesis, Massachusetts Institute of Technology.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2020. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 2493–2500.

Pan, M.; Huang, W.; Li, Y.; Zhou, X.; and Luo, J. 2020. xgail: Explainable generative adversarial imitation learning for explainable human decision analysis. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1334–1343.

Petsiuk, V.; Das, A.; and Saenko, K. 2018. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Scheller, C.; Schraner, Y.; and Vogel, M. 2020. Sample efficient reinforcement learning through learning from demonstrations in minecraft. In *NeurIPS 2019 Competition and Demonstration Track*, 67–76. PMLR.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shu, T.; Xiong, C.; and Socher, R. 2017. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. *arXiv preprint arXiv:1712.07294*.

Torabi, F.; Warnell, G.; and Stone, P. 2018. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*.

Xie, Y.; Vosoughi, S.; and Hassanpour, S. 2022. Towards Interpretable Deep Reinforcement Learning Models via Inverse Reinforcement Learning. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 5067–5074. IEEE.

Yu, T.; Finn, C.; Xie, A.; Dasari, S.; Zhang, T.; Abbeel, P.; and Levine, S. 2018. One-shot imitation from observing humans via domain-adaptive meta-learning. *arXiv preprint arXiv:1802.01557*.

Zhang, D.; Li, Q.; Zheng, Y.; Wei, L.; Zhang, D.; and Zhang, Z. 2021. Explainable hierarchical imitation learning for robotic drink pouring. *IEEE Transactions on Automation Science and Engineering*.

Zheng, B.; Verma, S.; Zhou, J.; Tsang, I. W.; and Chen, F. 2022. Imitation Learning: Progress, Taxonomies and Challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 1–16.