TAGIL: Temporal Attention Guided Imitation Learning

Luca von Mayer

Supervisor(s):
Steven James
Benjamin Rosman
2427051@students.wits.ac.za

Abstract

The Atari-HEAD dataset has been used extensively for many reinforcement learning tasks, notably for gaze and action prediction in imitation learning models. However, its temporal data (recorded reaction times) has been largely overlooked. In this work, we propose utilizing this temporal component—decision durations—to enhance imitation learning. Specifically, we build a model that predicts the difficulty of states, represented by human reaction times, assuming that longer reaction times indicate more challenging states. Alongside gaze predictions, we incorporate these difficulty predictions into an action prediction model, weighting difficult game states more heavily. We achieve this using a sample weighting method where harder states receive higher weights, influencing the loss function proportionally during training. Our results show that our approach improves action prediction across several Atari games.

`Code available at:` https://github.com/luca-vm/TAGIL

1 Introduction

Learning from human demonstration is a fundamental approach in the field of artificial intelligence, especially in applications such as robotics (Mandlekar et al. 2021), autonomous driving (Palazzi et al. 2018), and video games (Zhang et al. 2020). This technique leverages the wealth of knowledge embedded in human behavior, allowing agents to mimic complex decision-making processes that are often difficult to program explicitly. By studying how humans navigate environments and make decisions, we can train models that replicate this behavior more effectively, resulting in enhanced performance in tasks that demand high-level cognitive skills (Osa et al. 2018).

Much research has been conducted to address the challenges associated with learning from demonstration. Behavioral cloning trains agents through supervised learning using demonstrated actions and states as input (Bain and Sammut 1995). Attention Guided Imitation Learning (AGIL) enhances imitation learning by predicting human gaze locations from previous frames and applying this predicted attention through techniques like foveated rendering and attention masking (Zhang et al. 2018). Selective Eye-gaze Augmentation (SEA) selectively incorporates human gaze data during training, improving agent performance by determining when to use gaze information (Thammineni, Manjunatha, and Esfahani 2020). The Atari Human Eye-Tracking and Demonstrations (Atari-HEAD) dataset (Zhang et al. 2020), a collection of human gameplay data with corresponding eye-tracking and action information, has been extensively used for various reinforcement learning tasks and underpins many of these methodologies.

However, existing approaches, such as AGIL, suffer from notable limitations. They treat all states equally for gaze and action prediction, even though intuitively not all actions carry the same weight—certain moments require precise execution to avoid failure, while others are less critical. In autonomous driving, for example, reacting to a critical, time-sensitive decision, such as avoiding a potential accident, should be a major priority in learning. Similarly, in video games, difficult states often coincide with high-stakes challenges that demand more complex strategies. Understanding how humans take their time to approach harder states can greatly improve an agent's ability to replicate human decision-making.

This necessitates a model that differentiates between easy and hard states. Our key observation reveals we can utilize the temporal decision duration data available in the Atari-HEAD dataset as a proxy for state difficulty. Specifically, we predict reaction times, treating states that require longer human responses as inherently more challenging. By incorporating this data, we improve the learning process by focusing more on the difficult, high-stakes moments.

By combining gaze and difficulty predictions within a single model, we can output a state's difficulty along with its gaze map. We can then weigh hard states more heavily during the action prediction training process using a sample weighting method on our loss function. Our results demonstrate that this framework significantly enhances action prediction accuracy within the Atari domain, particularly in games like *Ms. Pac-Man*, *Berzerk*, and *Bank Heist*.

In the remainder of this paper, we first explore the background in Section 2, highlighting key insights from approaches such as AGIL. In Section 3, we describe the modifications we made to these existing methods, leading to the development of our temporal gaze and action prediction models. Section 4 presents our experimental results and a discussion of our findings. Section 5 describes other related work across this field. Lastly, in Section 6 we conclude with potential directions for future research.

2 Background

A Dataset for Imitation Learning

The Atari-HEAD dataset (Zhang et al. 2020), as described in Figure 1, provides a rich foundation for studying human gameplay behavior. Collected across 20 Atari games in a "semi-frame-by-frame" mode, it captures precise alignments between actions and gaze points by waiting at each frame for the human subject to input their action. This makes it particularly valuable for understanding human decision-making in game environments.

The dataset includes recordings from four subjects playing 16 games over 175 trials, totaling 2.97 million frames. It consists of image frames, human actions, reaction times, gaze positions, and rewards. The dataset is organized into:

1. **meta-data.csv:** This file contains metadata for the dataset, including game names, trial numbers, human subject identifiers, game start/load information, frame averaging status, frames per second, number of frames, eye-tracking validation error, and best scores obtained.
2. ***.tar.bz2 files:** These files contain game image frames, with each filename indicating its trial number.
3. ***.txt files:** These are label files for each trial, containing frame IDs, episode numbers, game scores, decision durations, unclipped rewards, actions, and gaze positions for each frame.

This comprehensive dataset has become instrumental in research aimed at understanding and replicating human gameplay behavior, such as training agents to predict saliency and improve imitation learning (Saran et al. 2021; Thammineni, Manjunatha, and Esfahani 2020).

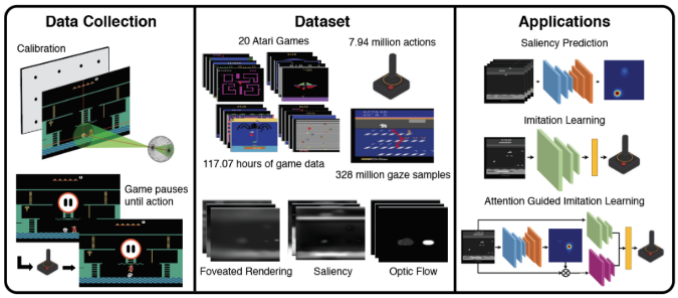


Figure 1: Overview of the Atari-HEAD dataset: data collection process, key data fields (e.g. image frames, actions, gaze points, reaction times), and research applications (Zhang et al. 2020).

Reinforcement Learning

To understand how we can leverage this human gameplay data, we first need to examine the fundamental framework of Reinforcement Learning (RL). RL provides the mathematical foundation for training agents to interact with an environment by making decisions that maximize cumulative rewards over time. This process can be mathematically represented as a Markov Decision Process (MDP), where at each time step t, the agent observes the current state s_t, takes action a_t, and transitions to the next state s_{t+1} depending on the environment's transition dynamics. The agent is then rewarded R_t based on the action taken and the resulting state.

The objective is to learn an optimal policy, π^*, that maps states to actions in a manner that maximizes the agent's expected long-term reward (Kaelbling, Littman, and Moore 1996). Mathematically, the optimal policy is defined as:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right],$$

where $\mathbb{E}$ represents the expectation over all possible outcomes, γ is a discount factor that weighs future rewards, and R_t is the reward at time step t. While this framework provides the foundation for decision-making, the challenge lies in determining how to effectively learn these optimal policies, particularly in complex environments.

Saliency and Visual Attention

One key insight from the Atari-HEAD dataset is the importance of visual attention in human decision-making. This brings us to the concept of saliency, which describes where humans focus their attention when making decisions.

A saliency map identifies the most visually prominent areas of an image, indicating where humans tend to focus their attention. Typically represented as a grayscale or heatmap, the intensity of each pixel indicates its saliency, with brighter areas being more attention-grabbing. These maps are key for understanding which parts of an image are most relevant, offering insights into human perception.

Early Saliency Prediction Saliency prediction can be divided into bottom-up and top-down approaches. Bottom-up methods, such as the Itti-Koch model (Itti, Koch, and Niebur 1998), generate saliency maps from low-level image features like color and orientation. Top-down models, in contrast, account for task relevance and goals.

Peters and Itti (2007) introduced a combined model incorporating bottom-up and top-down cues to predict human gaze during complex tasks. The bottom-up component generates a saliency map using 12 feature channels sensitive to color contrast, luminance, and motion energy, while the top-down component associates image features with likely gaze positions via a learned linear mapping, represented as:

$$W = F^+ \times P,$$

where W is the learned weight matrix, F^+ is the pseudo-inverse of the feature matrix, and P is the gaze density matrix. This understanding of visual attention has proven crucial for developing more sophisticated approaches to imitation learning.

Imitation Learning

Early Imitation Learning Building on our understanding of both reinforcement learning and visual attention, imitation learning provides a framework for leveraging human demonstration data to train effective agents. Behavioural

cloning (Bain and Sammut 1995) is a foundational approach whereby agents are trained via supervised learning using demonstrated actions and states as input. Behavioural cloning from observation (Torabi, Warnell, and Stone 2018) consists of an iterative two-phase method where the agent first obtains self-supervised experience by predicting the missing piece of given input. The agent then learns only from state observations, employing an inverse dynamics model to make up for missing actions. The inverse dynamics model, $M_\theta : S \times S \rightarrow p(A)$, maps state transitions $(s_t, s_{t+1}) \in T^\pi$ to a distribution of possible actions that could have caused the transition under policy π.

Attention Guided Imitation Learning (AGIL) Recognizing that visual attention plays a crucial role in human decision-making, Zhang et al. (2018) introduced AGIL, a model that predicts human gaze locations using deep neural networks, then applies this predicted attention through techniques like foveated rendering (reducing image quality in peripheral areas) and attention masking (using predicted heatmaps to highlight focus areas). AGIL improves action prediction accuracy and game performance compared to traditional imitation learning methods, laying the groundwork for gaze-based approaches used in later research.

3 TAGIL: Policy Network with Temporal and Visual Attention

This research presents a framework for improving imitation learning in Atari games by incorporating both visual attention and temporal information from human gameplay. The framework consists of two main components: a gaze prediction network and a policy network. We compare two approaches:

1. **Baseline Approach (AGIL):** Uses a standard gaze prediction network to generate saliency maps, which are then fed into the AGIL policy network to predict actions as per Zhang et al. (2018).
2. **Enhanced Approach (TAGIL):** Extends the baseline by incorporating temporal information through a modified gaze prediction network that also classifies state difficulty. This information is then used in an enhanced policy network that weighs difficult states more heavily during training.

Baseline

Gaze Prediction Network The baseline gaze prediction network follows the architecture and preprocessing techniques described in Zhang et al. (2018). This model treats gaze prediction as a saliency prediction task, aiming to generate a probability distribution (saliency map) that indicates likely gaze locations.

The ground truth gaze maps were generated by plotting the gaze points from the dataset. For each gaze point, a 17×17 pixel Gaussian blur was applied to create a smoothed region, rather than a sharp point. This approach encouraged the model to predict a more general area for gaze positions, rather than focusing on precise pixel locations. The Gaussian blur reflects the natural uncertainty in human gaze, and it provides a more realistic target for the network to learn.

The network consists of three branches:

1. **Image Frames:** The top branch processes game frames by stacking four consecutive grayscale frames, each resized to 84×84 pixels. Stacking frames helps capture temporal context, addressing the non-Markovian nature of single frames.
2. **Optical Flow:** The middle branch computes optical flow between consecutive frames to capture motion information, which is essential for predicting human attention in dynamic environments (Farnebäck 2003).
3. **Itti-Koch Saliency Map:** The bottom branch incorporates bottom-up saliency maps generated using the Itti-Koch model (Itti, Koch, and Niebur 1998), highlighting areas of visual interest based on low-level features.

The outputs from these branches are averaged to generate the saliency map. The model is trained using the Kullback-Leibler (KL) divergence loss function, defined as:

$$\mathrm{KL}(P, Q) = \sum_i Q_i \log \left(\epsilon + \frac{Q_i}{P_i + \epsilon} \right),$$

where P is the predicted saliency map, Q is the ground truth, and $\epsilon = 1 \times 10^{-10}$ is a constant for numerical stability.

AGIL The baseline model by Zhang et al. (2018) follows a two-channel deep neural network architecture. Raw image frames are processed in the top channel, and simultaneously, the bottom channel handles images that have been element-wise multiplied with predicted gaze saliency maps. This attention mask emphasizes regions attended by the human gaze.

The outputs of both channels are then averaged to produce a final action prediction. The architecture extends the Deep Q-Network (DQN) structure (Mnih et al. 2015), with the addition of visual attention information from predicted gaze heatmaps. The model is designed to imitate human actions by integrating visual attention into the decision-making process.

TAGIL: Temporal attention Guided Imitation Learning

Temporal Gaze (T-Gaze) Prediction Network The temporal gaze prediction network extends the baseline model to predict both the gaze heatmap and the difficulty of game states, classified as "easy" or "hard" based on reaction time. The original network architecture was preserved, but an additional difficulty prediction head was added to handle this binary classification task.

- **Difficulty Prediction Head:** An additional branch was introduced after the final concatenation of the outputs from the image frames, optical flow, and saliency map branches. This branch consists of several dense layers:
  - Two fully connected layers (128 and 64 units) with ReLU activations and L2 regularization.
  - Dropout layers added for regularization.

– A final output layer with a softmax activation, producing a binary classification for "easy" (reaction time < 60 ms) or "hard" (reaction time > 60 ms) states.

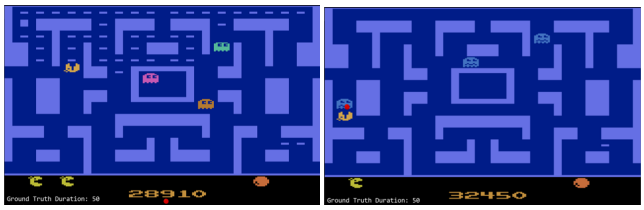
This head was designed to predict the difficulty class based on the concatenated features from all three branches.

Binary Classification for Difficulty: To facilitate a more straightforward approach to estimating task difficulty, reaction times were binned into two categories: "easy" for states with a reaction time < 60 ms and "hard" for those with > 60 ms. This proved more effective than regression due to the skewed distribution of reaction times—where hard states are very rare, representing about 0.1% of the data for certain games.

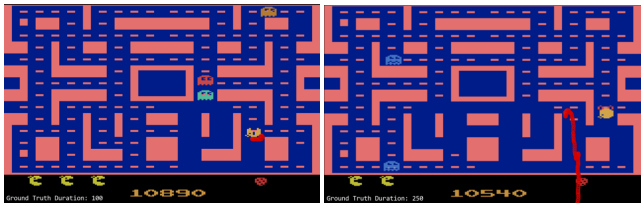
To illustrate the difference between easy and hard states, Figure 2 provides visual examples of *Ms. Pac-Man* states, marked with red dots indicating ground truth gaze predictions. These examples demonstrate how difficulty correlates with both visual complexity and reaction times.

In the first easy state (Figure 2a), the path forward is clear as the dots to consume are directly ahead, with no ghost threats nearby, making the decision to move upward an obvious one. Interestingly, the player's gaze isn't even focused on characters here, suggesting the simplicity of this state. In the second easy state (Figure 2b), Ms. Pac-Man has consumed a power-up, enabling her to chase ghosts, one of which is positioned directly above, prompting an easy decision to go up.

Conversely, the first hard state (Figure 2c) presents Ms. Pac-Man and nearby ghosts at a crossroads with multiple possible directions, creating ambiguity in determining the optimal path. In the second hard state (Figure 2d), Ms. Pac-Man has recently consumed a power-up, allowing her to chase ghosts. However, the ghosts are relatively far, requiring strategic planning to reach them efficiently.



(a) Easy State 1 (50ms) (b) Easy State 2 (50ms)



(c) Hard State 1 (100ms) (d) Hard State 2 (250ms)

Figure 2: Examples of *Ms. Pac-Man* states classified as easy and hard. Red dots indicate ground truth gaze predictions. Reaction times (ms), a proxy for difficulty, align with the intuitive difficulty of these states.

- **Handling Data Imbalance:** Due to the significant imbalance between easy and hard states in the ATARI-HEAD dataset, two techniques were employed:
  – **SMOTE:** Synthetic Minority Over-sampling Technique (Chawla et al. 2002) was used to generate synthetic data points for the minority class (hard states).
  – **Focal Loss:** Focal loss (Lin et al. 2018) was used to emphasize learning from the minority class as well as minimizing the impact of easier, well-classified states:

$$\text{FL}(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t),$$

  here p_t represents the predicted probability of the true class, α_t serves as a weighting factor, and γ is a focusing parameter that reduces the loss contribution from well-classified examples.
- **Weighted Loss Function:** The final loss function combined the output from the gaze prediction and difficulty prediction heads. After trial and error, a weighted loss was chosen, with 0.3 for gaze prediction and 0.7 for difficulty prediction, emphasizing the importance of correctly classifying hard states. Thus the final loss for this network is given by:

$$\text{CombinedLoss} = 0.3 \times KL(P,Q) + 0.7 \times FL(A,B),$$

  where P represents the predicted saliency map, and Q denotes the ground truth saliency map. A is the predicted difficulty and B is the ground truth difficulty. KL and FL are the Kullback-Leibler divergence and Focal losses defined earlier.

The overall architecture of the temporal gaze prediction network is illustrated in Figure 3.

TAGIL In TAGIL, we extend the baseline AGIL architecture by incorporating an additional input that represents the difficulty of each game state. This difficulty input is fed into the model alongside the visual features extracted from the image frames. The difficulty classification is concatenated with the flattened convolutional features before passing through the fully connected layers. By introducing this input, the model can account for the complexity of each state while predicting actions, allowing it to adjust its decision-making process based on the perceived difficulty.

The model employs a sample weighting mechanism where each state i is assigned a weight w_i based on its difficulty classification: hard states receive a weight of α (where $\alpha > 1$ is a hyperparameter), while easy states receive a weight of 1. For each individual training example, its contribution to the loss is scaled by its corresponding weight.

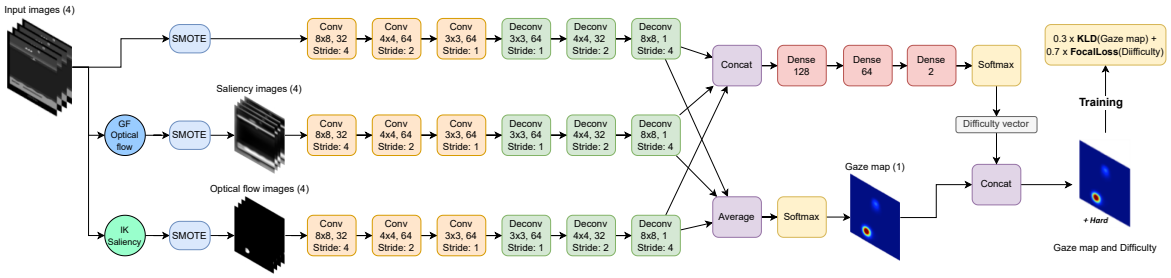


Figure 3: Temporal Gaze Network: Model shows the two prediction heads (gaze heatmap and difficulty classification) and the flow of features from the different input branches to the combined output. Lastly, it shows the loss function during training.

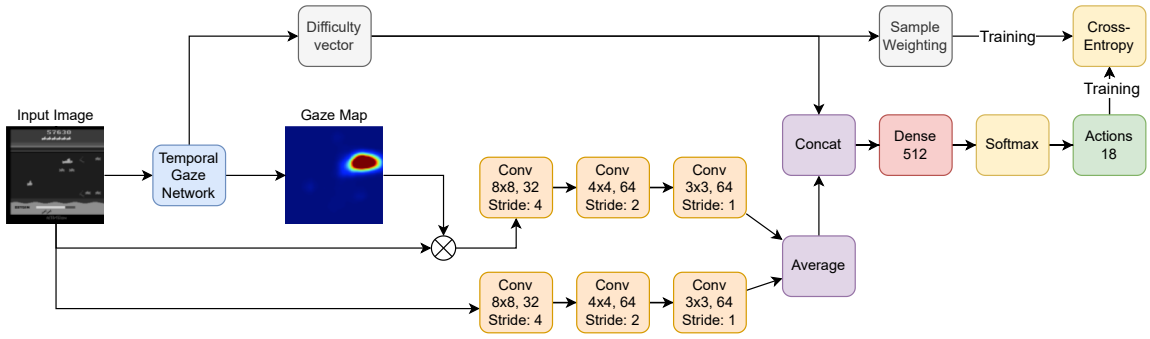


Figure 4: TAGIL: The model incorporates an additional input for difficulty classification, which is concatenated with the visual features extracted from image frames. A sample weighting mechanism emphasizes harder game states during training, allowing the model to prioritize challenging cases in its learning process.

The weighted sparse categorical cross-entropy loss for a batch of N samples is given by:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} w_i a_i \log(\hat{a}_i),$$

where:

- w_i is the sample weight (α for hard states, 1 for easy states)
- a_i is the true action label for sample i
- $\hat{a}_i$ is the predicted probability for the correct action

This weighted loss function ensures that during training, errors on hard states result in larger gradients (scaled by α), causing the model to make larger parameter updates when it makes mistakes on difficult states. The magnitude of this emphasis is controlled by the weighting factor α, which serves as a tunable hyperparameter.

The overall architecture, including the addition of the difficulty input and sample weighting mechanism, is shown in Figure 4.

4 Experiments

Temporal Gaze

In this section, we compare the performance of our T-Gaze model against the baseline in predicting human gaze points during gameplay. The primary metric used for evaluation is Intersection over Union (IoU), which measures the overlap between the predicted gaze maps and the ground truth gaze maps (with a threshold of 0.3 used for this binarization).

The gaze models were trained until early stopping was triggered for 10 runs per game on both the Gaze and T-Gaze models. Table 1 shows the IoU results of these runs on validation data.

Game	GAZE	T-GAZE
Ms. Pacman	**0.2494 ± 0.003**	0.2483 ± 0.003
Breakout	0.3220 ± 0.014	**0.3337 ± 0.002**
Freeway	**0.3939 ± 0.004**	0.3833 ± 0.002
Bank Heist	**0.3376 ± 0.005**	0.3354 ± 0.007
Montezuma's Revenge	**0.4012 ± 0.006**	0.3807 ± 0.003
Berzerk	**0.2845 ± 0.007**	0.2717 ± 0.001

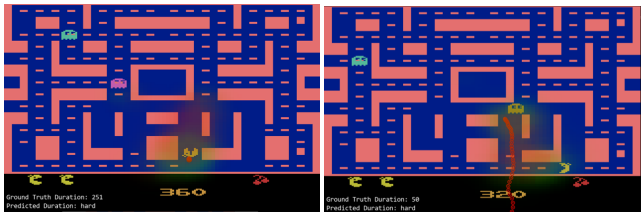
Table 1: Validation IOU (mean ± standard deviation) between Gaze and T-Gaze models across 6 Atari games.

Overall, the T-Gaze model performs slightly worse than the baseline across multiple games, as reflected by the IoU

on validation data. A key factor contributing to this performance drop is the use of synthetic data generated by SMOTE to produce hard states. These synthetic samples, created through linear interpolation, are not true game states and therefore lack the accuracy of real gameplay data. This inaccuracy makes it harder for the model to predict gaze points in artificially generated states, leading to a drop in precision. Additionally, for harder game states, human gaze patterns may inherently be less predictable due to the increased complexity or chaotic nature of the state, further complicating gaze prediction.

The amount of synthetic data generated is controlled by the SMOTE hyperparameter, "minority increase." Adjusting this parameter introduces a trade-off: while additional synthetic data improves the model's ability to recognize hard states, it also reduces gaze prediction accuracy as the model contends with these interpolated, less accurate representations of real gameplay. Future work could investigate optimizing synthetic data use to maintain high accuracy in gaze prediction while improving hard state identification.

In line with these findings, T-Gaze demonstrated effective learning in predicting state difficulty, as reflected by a steady reduction in focal loss across training and validation sets. The use of synthetic hard data significantly influenced predictions, with SMOTE-augmented data resulting in an increased frequency of hard-state predictions. While hard states initially comprised roughly 0.1% of the training data, post-training predictions identified as much as 20-30% of states as hard. Human players typically make an initial difficult decision and follow through on it across consecutive frames, even when those frames are visually similar. In comparison, T-Gaze frequently classified clusters of states as hard, identifying not just a single critical frame but a broader sequence requiring strategic decisions. This clustering reflects the intuitive nature of gameplay, where complex decisions often unfold over multiple frames rather than in isolation. Figure 5 provides an example comparison.



(a) (b)

Figure 5: Comparison between Gaze maps: showing baseline and T-Gaze heatmaps in green and red, respectively, with ground-truth gaze points as red dots, while the lower-left displays the difficulty predictions.

TAGIL Accuracy

In comparing the TAGIL model to the baseline AGIL model, both demonstrated effective learning with decreasing training and validation losses and increasing accuracies. Notably, TAGIL's training accuracy converged to a lower percentage than AGIL's, suggesting reduced overfitting tendencies. Validation accuracies were averaged over 10 runs, with early stopping applied if no improvement occurred for 40 epochs.

As shown in Table 2, TAGIL's performance varied significantly across different game types. TAGIL demonstrated superior performance in games characterized by complex navigation and dynamic decision-making, specifically *Ms. Pacman* (+8%), *Bank Heist* (+5%), and *Berzerk* (+10%). These games share common features including maze-like environments, multiple dynamic threats, and varying decision complexities. The incorporation of human reaction times appears particularly beneficial in these contexts, where temporal information helps distinguish between simple navigation and high-stakes decision points.

However, TAGIL showed comparable performance to AGIL in games with more structured gameplay patterns: *Montezuma's Revenge* (-1.0%), *Breakout* (+0.6%), and *Freeway* (-0.5%). Games like *Breakout* and *Freeway* typically feature more predictable patterns, constrained movement options, and consistent reaction timing requirements. While *Montezuma's Revenge* differs with its complex platforming mechanics and sparse rewards, its long-term planning requirements may limit the utility of short-term reaction time information. In these scenarios, the additional temporal information provided minimal advantages over the baseline visual features alone.

Game	AGIL	TAGIL
Ms. Pacman*	69.24 ± 0.5	$\mathbf{77.46 \pm 0.1}$
Breakout	83.79 ± 0.3	$\mathbf{84.45 \pm 0.3}$
Freeway	$\mathbf{93.28 \pm 0.3}$	92.85 ± 0.5
Bank Heist*	58.13 ± 0.4	$\mathbf{63.52 \pm 0.3}$
Montezuma's Revenge	$\mathbf{88.96 \pm 0.2}$	87.90 ± 0.5
Berzerk*	54.48 ± 0.8	$\mathbf{64.83 \pm 0.3}$

Table 2: Percentage validation accuracy (mean $\pm$ standard deviation) between AGIL and TAGIL models across 6 Atari games. The * indicates a statistically significant difference via t-test.

These results reveal a clear pattern: temporal augmentation provides the greatest benefit in games with dynamic decision spaces and complex threat assessment. In more deterministic environments with predictable action sequences, the baseline visual features appear sufficient for effective action prediction. This distinction suggests that human reaction times serve as a meaningful proxy for identifying crucial decision points in complex gameplay scenarios.

5 Related Work

The concept of determining easy versus hard states is crucial in improving learning efficiency and aligning agent behavior with human decision making. The TAGIL framework uniquely incorporates human reaction times as a proxy for state difficulty, providing an additional temporal lens through which to differentiate easy and hard decision points. In contrast, several related works have explored alternate approaches to prioritize learning on more challenging states.

Prioritizing State Difficulty in Learning Prioritized experience replay (Schaul 2015) introduces an approach where transitions are sampled based on their temporal difference (TD) error, which reflects the model's uncertainty in predicting them. States with higher TD-error are considered "hard" and are prioritized during training because they represent situations where the model struggles to make correct predictions. This method helps focus learning on challenging states, leading to faster convergence and better performance. By allocating more training time to difficult states, the model learns more efficiently, similar to how identifying and focusing on hard states can optimize action prediction in various domains.

Prioritized Sweeping Moore and Atkeson (1993) introduces prioritized sweeping, an algorithm that focuses computational effort on the most "interesting" parts of a Markov system, rather than treating all states equally. By maintaining a priority queue and updating the estimates for the highest priority states first, prioritized sweeping can quickly learn accurate predictions or optimal policies in areas of the state space that provide the most valuable information, even in large, complex systems. This demonstrates the benefits of selectively updating "easy" or "useful" states to guide the learning process, rather than treating all states as equally important. These approaches to prioritizing state difficulty are directly relevant to the current work, as they provide a foundation for identifying and leveraging hard states to enhance imitation learning.

Selective Eye-gaze Augmentation (SEA)

Building on the insights from prioritizing state difficulty, Thammineni, Manjunatha, and Esfahani (2020) introduced Selective Eye-gaze Augmentation (SEA) to enhance Atari game agents' performance in imitation learning by integrating human gaze data selectively. The SEA network consists of a gating network, a gaze prediction network, and an action prediction network, which work in tandem to decide when and how gaze data should influence the agent's action predictions. By selectively using human gaze data, SEA improves action prediction accuracy even with gaze data that may be noisy or irrelevant. This selective use of gaze data aligns with the idea of prioritizing learning on challenging states.

Visual identification of subgoals

The work by Palazzi et al. (2018) on predicting where drivers focus their attention using deep neural networks has implications beyond the video game domain. Their analysis of the DR(eye)VE dataset revealed that, while driving, humans tend to focus on the road's vanishing point, which may be considered analogous to a subgoal or intermediate destination within a game world. This suggests that the concept of identifying salient or critical states is not limited to video games, but can also be applied to real-world decision-making tasks, such as autonomous driving.

6 Conclusion and Future Work

In this work, we have presented TAGIL, a framework for improving imitation learning in Atari games by incorporating both visual attention and temporal information from human gameplay. Our approach extends the existing AGIL model by predicting the difficulty of game states, represented by human reaction times, and using this information to selectively weight hard states during the training of the action prediction model. The experiments conducted on several Atari games demonstrate the effectiveness of our approach. TAGIL outperformed the baseline AGIL model in games characterized by complex navigation and dynamic decision-making, such as *Ms. Pac-Man*, *Bank Heist*, and *Berzerk*. This suggests that leveraging temporal information, as a proxy for state difficulty, can significantly enhance an agent's ability to mimic human decision-making in challenging environments. However, in games with more structured and predictable gameplay patterns, the additional temporal information provided minimal advantages over the baseline visual features alone. This highlights the importance of understanding the nature of the task and the characteristics of the environment when determining the appropriate modeling approach.

A key direction for future work is to expand the domain of this topic beyond video games. The concept of identifying easy versus hard states using reaction times in the current context was enabled by the semi-frame-by-frame data collection, where humans had unlimited time to make decisions. This may not be representative of how easy and hard states are encountered in the real world. For instance, in the context of autonomous driving, a fast reaction time may indicate either a very easy state that could be reacted to immediately, or a very hard state that required a rapid response (e.g. a potential accident). Applying the TAGIL framework to data collected in real-world environments, such as self-driving cars, would be a valuable avenue for future research. This would involve modifying the model to differentiate between fast reactions triggered by easy states versus those triggered by hard states.

Additionally, further exploration of the hypothesis that the TAGIL model performs well on maze-like games but not on other genres would be worthwhile. This could be investigated by testing the model on a variety of game types and contexts to better understand the limitations and generalizability of the approach.

Overall, this work demonstrates that incorporating temporal attention and state difficulty into imitation learning is a valuable approach for developing agents that can emulate human decision-making in complex, dynamic environments. Continued exploration of these techniques holds promise for advancing the field of artificial intelligence and creating agents that can seamlessly interact with and learn from human demonstration.

7 Acknowledgments

The author wishes to extend heartfelt thanks to their supervisors, Steven James and Benjamin Rosman, for their exceptional support and guidance throughout the research process.

References

Bain, M.; and Sammut, C. 1995. A Framework for Behavioural Cloning. In *Machine Intelligence 15*, 103–129.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.

Farnebäck, G. 2003. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, 363–370. Springer.

Itti, L.; Koch, C.; and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11): 1254–1259.

Kaelbling, L. P.; Littman, M. L.; and Moore, A. W. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4: 237–285.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2018. Focal Loss for Dense Object Detection. arXiv:1708.02002.

Mandlekar, A.; Xu, D.; Wong, J.; Nasiriany, S.; Wang, C.; Kulkarni, R.; Fei-Fei, L.; Savarese, S.; Zhu, Y.; and Martín-Martín, R. 2021. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533.

Moore, A. W.; and Atkeson, C. G. 1993. Prioritized sweeping: Reinforcement learning with less data and less time. *Machine learning*, 13: 103–130.

Osa, T.; Pajarinen, J.; Neumann, G.; Bagnell, J. A.; Abbeel, P.; Peters, J.; et al. 2018. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2): 1–179.

Palazzi, A.; Abati, D.; Solera, F.; Cucchiara, R.; et al. 2018. Predicting the Driver's Focus of Attention: the DR (eye) VE Project. *IEEE transactions on pattern analysis and machine intelligence*, 41(7): 1720–1733.

Peters, R. J.; and Itti, L. 2007. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *2007 IEEE conference on computer vision and pattern recognition*, 1–8. IEEE.

Saran, A.; Zhang, R.; Short, E. S.; and Niekum, S. 2021. Efficiently Guiding Imitation Learning Agents with Human Gaze. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 1109–1117.

Schaul, T. 2015. Prioritized Experience Replay. *arXiv preprint arXiv:1511.05952*.

Thammineni, C.; Manjunatha, H.; and Esfahani, E. T. 2020. Selective Eye-gaze Augmentation To Enhance Imitation Learning In Atari Games. *arXiv preprint arXiv:2012.03145*.

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

Zhang, R.; Liu, Z.; Zhang, L.; Whritner, J. A.; Muller, K. S.; Hayhoe, M. M.; and Ballard, D. H. 2018. AGIL: Learning attention from human for visuomotor tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 663–679.

Zhang, R.; Walshe, C.; Liu, Z.; Guan, L.; Muller, K.; Whritner, J.; Zhang, L.; Hayhoe, M.; and Ballard, D. 2020. Atari-HEAD: Atari human eye-tracking and demonstration dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 6811–6820.