Declarations

Funding

No funds, grants, or other support was received.

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval

Not applicable.

Consent to participate / Consent for publication

Not applicable.

Data availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Materials and/or Code availability

The source code is available at https://github.com/luca-vm/TAGIL. The repository includes the full pipeline (data processing, model training, and evaluation) and usage instructions. An archival snapshot is deposited on Zenodo with a DOI (https://doi.org/10.5281/zenodo.16756242).

Authors' contributions

L.V.M. designed the study, implemented the models, ran the experiments, analysed the data, and drafted the manuscript. S.J. and B.R. contributed to the study conception and methodology, provided supervision, aided interpretation of results, and revised the manuscript critically for important intellectual content. S.J. additionally assisted with writing the final version. All authors approved the final version and agree to be accountable for all aspects of the work.

Difficulty-Aware Imitation Learning with Temporal Gaze Guidance

Luca von Mayer[1], Benjamin Rosman[1,2], Steven James[1,2*]

[1]School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, 2000, South Africa.
[2]Machine Intelligence and Neural Discovery (MIND) Institute.

*Corresponding author(s). E-mail(s): steven.james@wits.ac.za;
Contributing authors: lucavonmayer@gmail.com;
benjamin.rosman1@wits.ac.za;

Abstract

Human reaction times provide a natural signal for estimating the difficulty of decision-making in sequential tasks: longer reaction times often correspond to more challenging states. In this work, we exploit this relationship to improve imitation learning by using reaction time as a proxy for state difficulty. We first train a model to predict state difficulty from reaction durations, alongside gaze predictions, and then incorporate these difficulty estimates into an action prediction model by weighting training samples to place greater emphasis on harder states. We evaluate our method on several Atari games using the Atari-HEAD dataset and show that incorporating difficulty weighting improves action prediction performance compared to existing gaze-based imitation learning approaches.

Keywords: imitation learning, gaze prediction

1 Introduction

Learning from human demonstration is a fundamental approach in artificial intelligence, with applications in domains such as robotics (Mandlekar et al. 2022), autonomous driving (Palazzi et al. 2018), and video games (Zhang et al. 2020). By leveraging the knowledge embedded in human behaviour, agents can be trained to reproduce complex decision-making processes that are otherwise difficult to specify explicitly. Observing how humans navigate environments and make decisions enables the development of models that replicate these behaviours, leading to improved performance on tasks requiring high-level cognitive skills (Osa et al. 2018).

A substantial body of work has addressed the challenges associated with learning from demonstration. Behavioural cloning trains agents via supervised learning, using demonstrated actions and states as input (Bain and Sammut 1995). Attention-Guided Imitation Learning (AGIL) improves imitation learning by predicting human gaze locations from previous frames and applying this predicted attention through techniques such as 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 identifying when gaze information is most useful (Thammineni et al. 2020). These methods have been widely evaluated using the Atari Human Eye-Tracking and Demonstrations (Atari-HEAD) dataset (Zhang et al. 2020), which pairs human gameplay with corresponding gaze and action information.

Despite these advances, existing approaches typically treat all states as equally important when predicting gaze and actions. In practice, not all decisions are of equal consequence. Some states demand precise, time-sensitive execution to avoid failure, while others allow for less precise behaviour. For example, in autonomous driving, reacting to an imminent collision requires far greater urgency and precision than following a vehicle on an open road. Similarly, in video games, high-stakes moments often demand more complex strategies and precise timing. We argue that capturing how humans slow down when approaching such states can significantly enhance an agent's ability to replicate human decision-making.

This motivates the need for a model that can distinguish between easy and difficult states. Our key insight is that the temporal decision duration data available in the Atari-HEAD dataset can serve as a proxy for state difficulty. Specifically, we predict reaction times, assuming that states requiring longer human responses are inherently more challenging. By incorporating this information, we guide the learning process to focus more heavily on difficult, high-stakes moments. By combining gaze and difficulty predictions within a single model, we produce both a gaze map and a difficulty estimate for each state. These difficulty scores are then used to weigh samples during action prediction training, such that harder states contribute more significantly to the loss function. Our results show that our approach improves action prediction accuracy across multiple Atari games, particularly in games like *Ms. Pac-Man*, *Berzerk*, and *Bank Heist*.

The remainder of this paper is structured as follows. In Section 2, we review the relevant background, highlighting key insights from approaches such as AGIL. Section 3 outlines our extensions to these methods, culminating in the development of our temporal gaze and action prediction models. In Section 4, we present our experimental results and discuss their implications. Section 5 surveys additional related work, and Section 6 concludes with potential directions for future research.

2 Background

Preliminaries

Reinforcement Learning (RL) provides a foundational framework for training agents to make decisions by interacting with an environment. At its core, RL aims to learn behaviours that maximise cumulative rewards over time, typically modelled as a Markov decision process

(MDP). An MDP consists of states, actions, transition dynamics, and rewards. At each time step t, the agent observes the current state S_t, selects an action A_t, and transitions to a new state S_{t+1} based on the environment's dynamics $\Pr(S_{t+1}|S_t, A_t)$ and receives reward R_t.

The objective is to learn an optimal policy, π^* that maps states to actions in a manner that maximises the expected cumulative reward over time (Kaelbling et al. 1996). Formally, the objective can be expressed as:

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

where $\mathbb{E}$ denotes the expectation over possible outcomes and $\gamma \in [0, 1)$ is a discount factor that prioritises immediate rewards over distant ones. While this formulation provides a principled basis for decision-making, solving RL problems in practice often requires large amounts of data and computational resources—particularly in high-dimensional environments.

Imitation Learning

Imitation learning offers an alternative approach to training agents by learning directly from expert demonstrations, thereby bypassing the need for extensive trial-and-error exploration. The most straightforward approach is behavioural cloning (Bain and Sammut 1995), where the problem is reduced to an instance of supervised learning. In this setting, we train a policy π to predict actions based on observed state-action pairs $\{(S_i, A_i)\}_{i=1}^N$ collected from expert behavior. The training objective can be formulated as

$$\min_{\pi} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\big(a_i, \pi(s_i)\big),$$

where $\mathcal{L}$ is a suitable loss function—often the cross-entropy loss for discrete actions or mean squared error for continuous actions. This formulation seeks to minimise the discrepancy between the actions taken by the expert and those predicted by the policy, thereby encouraging the model to replicate the expert's decision-making process.

In settings where expert actions are not available, behavioural cloning from observation (BCO) (Torabi et al. 2018) extends this framework by allowing the agent to learn purely from state trajectories. BCO proceeds in two phases: the agent first uses self-supervised learning to fill in the missing action information, and then trains a policy based on the reconstructed data. A key component of this approach is the inverse dynamics model $M_\theta : S \times S \rightarrow p(A)$, which predicts the distribution of actions that could have caused a transition from state S_t to state S_{t+1}. This enables agents to learn from demonstrations even when only state observations are available.

Visual Attention and Saliency

One key insight in human decision-making is the importance of visual attention. Saliency, which refers to where humans focus their attention during decision-making, plays a crucial role in how actions are chosen. A saliency map highlights the most visually prominent regions of an image, offering a pixel-wise representation of attentional focus. Brighter areas on the map indicate higher saliency—regions likely to draw human attention. These maps provide insight into which parts of a scene are most perceptually relevant, making them a valuable tool for modelling human perception and guiding agent behaviour.

Early Saliency Prediction

Saliency prediction methods are typically classified into bottom-up and top-down approaches. Bottom-up methods, such as the Itti-Koch model (Itti et al. 1998), rely on low-level visual features like colour, luminance, and orientation to compute saliency maps. These methods are task-agnostic, capturing stimulus-driven attention. Top-down approaches, by contrast, incorporate knowledge about the observer's goals or tasks, adjusting the saliency map according to the task the individual is performing.

Peters and Itti (2007) introduce 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.

Attention-Guided Imitation Learning (AGIL)

Building on these insights, Zhang et al. (2018) propose Attention-Guided Imitation Learning (AGIL), a framework that integrates visual attention into the imitation learning pipeline. AGIL first predicts human gaze using a deep neural network trained on gameplay footage. The predicted saliency maps are then used to guide the agent's perception through two mechanisms: foveated rendering, which reduces resolution in peripheral regions of the image, and attention masking, which highlights regions of interest using the predicted saliency heatmaps.

AGIL significantly improves the accuracy of action prediction and overall agent performance. By incorporating human-like visual attention, AGIL outperforms traditional imitation learning approaches and provides a foundation for subsequent work that integrates gaze into policy learning.

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 et al. 2020).

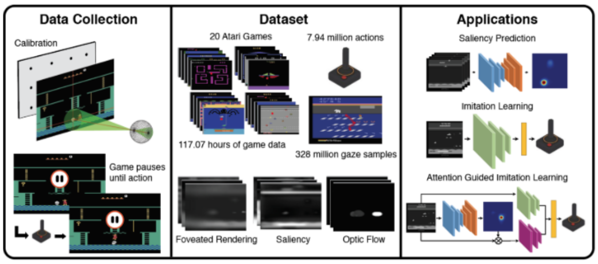


Fig. 1: Overview of the Atari-HEAD dataset: data collection process, key data fields (e.g. image frames, actions, gaze points, reaction times), and research applications (Zhang et al. 2020).

3 Imitation Learning with Decision Times

Imitation learning methods that incorporate human gaze typically treat all states as equally important, even though human behaviour reveals that some states are more demanding than others. Reaction time offers a simple and observable proxy for decision difficulty: longer reaction times often correspond to states that are harder to resolve. In this work, we propose a method that uses reaction time to guide both representation learning and policy learning. Specifically, we aim to predict *temporal* gaze to capture how attention evolves over short frame sequences, and to predict a per-state *difficulty* label from the same features. The predicted gaze will be used to modulate the visual input to the policy, while the predicted difficulty will be used to emphasise harder states during training via sample weighting. This approach is intended to focus learning on the states that are most critical for task success.

An overview of our architecture is shown in Figure 2. *Temporal Attention-Guided Imitation Learning* (TAGIL) extends the baseline AGIL framework by incorporating temporal information into gaze prediction and by weighting policy learning according to predicted difficulty. TAGIL consists of three main components: a temporal gaze prediction network that uses multiple consecutive frames to capture motion and temporal dependencies; a difficulty classifier that predicts whether a state is “easy” or “hard” using reaction time as a proxy for difficulty; and a policy network with difficulty weighting that uses predicted gaze to modulate the visual input and predicted difficulty to weight the action-prediction loss. Each of these components is described in detail in the following subsections.

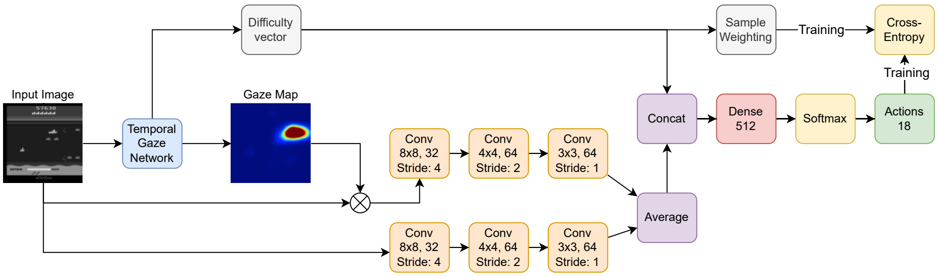


Fig. 2: **TAGIL architecture**. The framework consists of a temporal gaze prediction network that captures motion cues from consecutive frames, and a difficulty classifier that estimates state difficulty from reaction time. The predicted difficulty is concatenated with visual features before action prediction, and a sample-weighting mechanism increases the influence of harder states during training.

Gaze and Action Prediction in AGIL

Our approach builds on the *Attention-Guided Imitation Learning* (AGIL) framework (Zhang et al. 2018), which predicts a saliency map of likely human gaze locations and incorporates it into an action prediction network to imitate demonstrator behaviour.

AGIL formulates gaze prediction as a saliency estimation problem. Ground-truth gaze maps are generated from the dataset’s recorded gaze points by applying a 17×17 pixel Gaussian blur to each point, creating smooth regions rather than sharp points. This encourages the model to predict broader regions of interest, reflecting the natural uncertainty of human gaze and providing a more realistic target for learning.

The gaze network processes input through three parallel branches:

1. **Image frames**: Four consecutive grayscale game frames, each resized to 84×84 pixels, are stacked along the channel dimension. This captures short-term temporal context and mitigates the non-Markovian nature of single frames.
2. **Optical flow**: Dense optical flow is computed between consecutive frames using the Farnebäck method (Farnebäck 2003), providing motion cues that are often critical for predicting human attention in dynamic scenes.
3. **Saliency map**: Bottom-up saliency maps are generated using the Itti-Koch model (Itti et al. 1998), highlighting visually distinctive areas based on low-level features such as colour, intensity, and orientation.

The outputs from these three branches are averaged to produce the predicted saliency map. The network is trained using the Kullback-Leibler (KL) divergence:

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

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

The predicted saliency map is integrated into the action prediction stage through a two-channel deep neural network. One channel processes the raw image frames, while the other processes the same frames after element-wise multiplication with the predicted saliency map, producing an attention-masked representation that emphasises gaze-attended regions.

The outputs of the two channels are averaged to produce the final action prediction. Architecturally, AGIL extends the Deep Q-Network (DQN) design (Mnih et al. 2015) by incorporating visual attention from predicted gaze heatmaps, enabling the policy to imitate human actions while explicitly attending to regions deemed important by human demonstrators.

Temporal Gaze Prediction Network

The temporal gaze (T-Gaze) prediction network extends the gaze model above to produce two outputs: a gaze heatmap and a binary difficulty classification for each game state. The overall architecture is shown in Figure 3. The base network processes three input streams—image frames, optical flow, and Itti–Koch saliency maps—whose features are concatenated before branching into the two prediction heads.

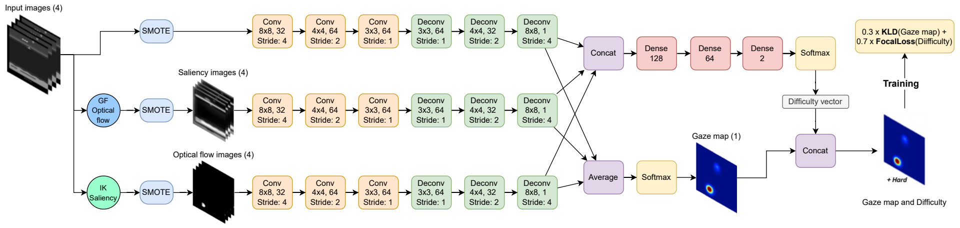
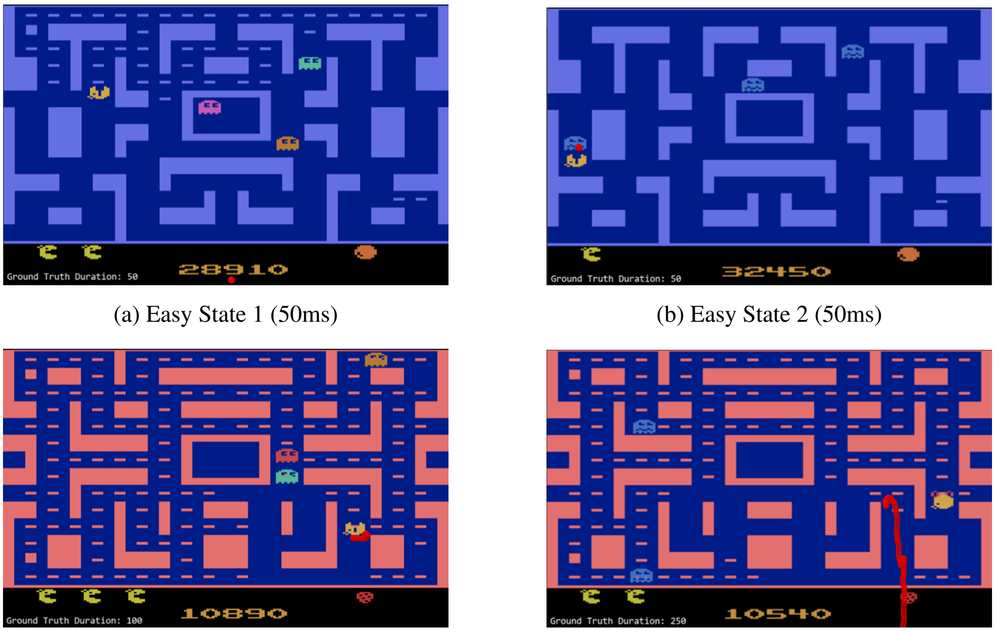


Fig. 3: **Temporal Gaze Network**. Two prediction heads (gaze heatmap and difficulty classification) operate on features from three input branches: image frames, optical flow, and saliency maps. The combined loss function balances saliency alignment and difficulty classification.

An additional difficulty prediction head is introduced after the final feature concatenation. This head consists of two fully connected layers (128 and 64 units) with ReLU activations and L2 regularisation, interleaved with dropout layers, followed by a softmax output producing the probability of each difficulty class.

Difficulty labels are derived from reaction time statistics in the Atari-HEAD dataset: states with a reaction time less than 60 ms are labelled *easy*, and those with greater than 60 ms are labelled *hard*. This thresholded classification approach is preferred to regression due to the skewed distribution of reaction times, where hard states can comprise as little as 0.1% of the data in some games.

To illustrate the difference between easy and hard states, Figure 4 shows examples of *Ms. Pac-Man* states with red dots indicating ground-truth gaze locations. In the first easy state (Figure 4a), the path is clear and ghosts are absent, making the upward move obvious; in the second (Figure 4b), a nearby ghost can be chased immediately after a power-up. By contrast, the first hard state (Figure 4c) presents multiple ghosts and route options, while the second (Figure 4d) requires strategic planning to reach distant targets.


Ground Truth Duration: 50
28910
Ground Truth Duration: 50
32450
Ground Truth Duration: 100
10890
Ground Truth Duration: 250
10540

(a) Easy State 1 (50ms)

(b) Easy State 2 (50ms)

(c) Hard State 1 (100ms)

(d) Hard State 2 (250ms)

Fig. 4: Examples of *Ms. Pac-Man* states classified as easy and hard. Red dots indicate ground-truth gaze. Reaction time (ms) serves as a proxy for difficulty and aligns with intuitive task complexity.

The severe imbalance between easy and hard states is addressed using two strategies. First, the *Synthetic Minority Over-sampling Technique* (SMOTE) (Chawla et al. 2002) generates synthetic minority-class samples by interpolating between existing feature vectors. Second, *focal loss* (Lin et al. 2017) emphasises hard examples and reduces the loss contribution from well-classified easy states. The focal loss between predicted probabilities A and ground-truth labels B is defined as:

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

where p_t is the predicted probability assigned to the true class given A and B, α_t is a class weight, and γ is a focusing parameter.

The T-Gaze network is trained jointly on gaze prediction and difficulty classification. The gaze branch is optimised with the Kullback–Leibler divergence $\mathrm{KL}(P, Q)$ between predicted and ground-truth saliency maps. The difficulty branch is trained with focal loss $\mathrm{FL}(A, B)$ between predicted and ground-truth difficulty labels.

The combined objective is:

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

where the loss weights were selected empirically to emphasise accurate classification of hard states.

Temporal Attention-Guided Imitation Learning

In TAGIL, we extend the baseline AGIL architecture to incorporate the predicted difficulty of each game state as an additional input to the policy. The difficulty score from the T-Gaze classifier is concatenated with the flattened convolutional features before passing through the fully connected layers that produce the action prediction. This allows the policy to adapt its decision-making process according to the estimated complexity of the current state.

To prioritise learning from challenging states, we apply a sample weighting mechanism. Each training example i is assigned a weight w_i: hard states receive $w_i = \alpha$ (where $\alpha > 1$ is a hyperparameter) and easy states receive $w_i = 1$. The weighted sparse categorical cross-entropy loss for a batch of N samples is:

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

where a_i is the one-hot encoded ground-truth action for sample i and $\hat{a}_i$ is the predicted probability of that action. This weighting increases the gradient magnitude for mistakes on hard states by a factor of α, leading the model to focus updates on these cases. The value of α is tuned to balance emphasis on difficult states with coverage of easier ones.

4 Experiments

In this section, we evaluate our approach against the baseline AGIL framework. We consider two aspects of performance: (i) gaze prediction accuracy, to determine whether the addition of difficulty prediction compromises visual attention modelling, and (ii) imitation learning accuracy, which is the primary objective of our approach. All experiments are conducted on six Atari games, chosen to cover a mixture of reactive and deliberative decision-making styles, with results averaged over 10 runs.

Gaze Accuracy

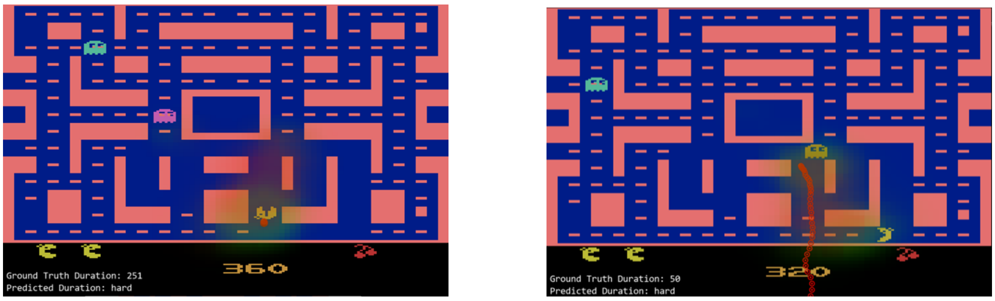
We first compare our T-Gaze model to the baseline gaze network in predicting human gaze points during gameplay. Accuracy is measured using Intersection over Union (IoU) between predicted and ground-truth gaze maps, binarised at a threshold of 0.3. Table 1 reports the mean and standard deviation over 10 runs.

Game	AGIL Gaze Network	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 the original gaze network and T-GAZE model across six Atari games.

The results show that T-Gaze performs slightly below the baseline in most games, with the exception of *Breakout*. These differences are small, indicating that the addition of difficulty prediction has minimal impact on gaze accuracy. Since gaze prediction is not the end goal, and T-Gaze remains competitive, any small drop in this metric is acceptable, provided action prediction performance improves.

The slight reduction in IoU is linked to the use of SMOTE-generated hard states, which are created by interpolating between real gameplay frames. While these synthetic samples improve the model's ability to recognise hard states, they do not perfectly capture the spatial and temporal patterns of true states, making gaze locations less precise. The proportion of synthetic data is controlled by the SMOTE "minority increase" hyperparameter, which introduces a trade-off between difficulty recognition and gaze accuracy. Post-training, we observed that the model identified 20–30% of frames as hard (compared to the original 0.1%), often marking clusters of consecutive frames—a behaviour consistent with the fact that complex decisions in gameplay typically unfold over multiple steps. Figure 5 shows an example comparison of baseline and T-Gaze gaze maps with corresponding difficulty predictions.



(a) (b)

Fig. 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

Imitation Learning Accuracy

We next compare TAGIL and AGIL on the action prediction task, which is the primary objective of our approach. Validation accuracies were averaged over 10 independent runs, with early stopping applied if no improvement was observed for 40 epochs.

As shown in Table 2, TAGIL's performance varied by game type. It achieved substantial and statistically significant gains in *Ms. Pacman* (+8%), *Bank Heist* (+5%), and *Berzerk* (+10%). These titles share characteristics such as maze-like layouts, multiple dynamic threats, and decision points with varying complexity. In such environments, the incorporation of human reaction time as a difficulty signal appears particularly beneficial, enabling the model to distinguish between routine navigation and high-stakes decision moments.

In contrast, performance was comparable to AGIL in *Breakout* (+0.6%), *Freeway* (–0.5%), and *Montezuma's Revenge* (–1.0%). *Breakout* and *Freeway* involve highly predictable gameplay patterns, limited movement options, and consistent reaction timing, reducing the added value of difficulty weighting. *Montezuma's Revenge* has complex platforming and sparse rewards, but its long-term planning requirements limit the usefulness of short-term reaction-time cues. In these cases, the baseline's visual features alone suffice for high performance.

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: Validation accuracy (%, mean ± standard deviation) for AGIL and TAGIL across six Atari games. The * indicates a statistically significant difference via t-test.

These results reveal a consistent pattern: temporal augmentation offers the greatest benefit in games with dynamic decision spaces and complex threat assessment, where difficulty weighting highlights high-stakes moments. In more deterministic environments with predictable action sequences, the baseline visual features appear sufficient for accurate action prediction. Overall, this supports the view that human reaction time is an effective proxy for identifying crucial decision points in complex gameplay scenarios.

5 Related Work

A central idea in this work is that learning can be made more efficient by focusing on the most informative or challenging parts of the state space. TAGIL applies this principle by using human reaction time as a proxy for state difficulty, allowing the model to emphasise high-stakes situations during imitation learning. This connects to a broader body of work on prioritising state difficulty, selectively incorporating attention signals, and identifying salient decision points.

Prioritising state difficulty in learning

Several reinforcement learning approaches have explored ways to concentrate updates on difficult or valuable states. Prioritised experience replay (Schaul 2015) samples transitions according to their temporal-difference error, giving preference to states where the agent's predictions are most uncertain. Prioritised sweeping (Moore and Atkeson 1993) maintains a queue of high-priority states and updates them first, directing computation to the most informative regions of the state space. Curriculum and self-paced learning methods extend this idea by organising training from easy to hard or adapting the pace based on the learner's progress (Klink et al. 2020; Liu et al. 2021). In imitation learning, difficulty-aware weighting has been applied to filter or down-weight imperfect demonstrations (Wang et al. 2021). TAGIL differs from these agent-centric approaches by deriving its difficulty signal directly from human behaviour.

Selective use of gaze information

Human gaze provides a rich source of task-relevant information, but not all gaze data is equally useful. Thammineni et al. (2020) address this by learning a gating mechanism that determines when gaze information should influence action prediction, improving performance by avoiding noisy or irrelevant gaze cues. Other gaze-guided imitation learning frameworks include GRIL (Thakur et al. 2021), which jointly predicts gaze and actions via multi-objective learning; coverage-based gaze loss (Saran et al. 2021), which uses gaze to guide attention in behaviour cloning; and gaze-based regularisation (Banayeeanzade et al. 2025), which mitigates causal confusion by encouraging attention to causally relevant features in Atari and CARLA driving tasks. In robotics, gaze has been integrated with foveated vision to improve manipulation efficiency and precision (Kim et al. 2021), and extended to memory-based gaze prediction for tasks requiring recall of previously seen objects (Kim et al. 2022).

Identifying salient or critical states

Beyond games, research in other domains has examined how attention patterns align with decision-making priorities. In driving, Palazzi et al. (2018) found that drivers often focus on the road's vanishing point: a visual subgoal guiding navigation. In human-robot interaction, gaze collected via natural demonstration interfaces has been shown to yield more transferable and task-relevant patterns than restricted-view devices (Ishida et al. 2025). Foveated vision models have also been shown to reduce computational requirements while preserving performance in control tasks (Chen et al. 2025). These findings reinforce the view that gaze can act as a proxy for identifying subgoals and high-stakes moments. TAGIL builds on this idea by linking spatial attention with a temporal measure—reaction time—to identify and emphasise demanding states during imitation learning.

6 Conclusion and Future Work

In this work, we introduced a framework for improving imitation learning in Atari games by combining visual attention with temporal information from human gameplay. Our approach extends the AGIL model by predicting the difficulty of game states, using human reaction times as a proxy, and weighting harder states more heavily during action prediction training.

Experiments across six Atari games showed that TAGIL outperforms AGIL in environments with complex navigation and dynamic decision-making, such as *Ms. Pacman*, *Bank Heist*, and *Berzerk*, where temporal information helps distinguish routine actions from high-stakes decisions. In contrast, for games with highly structured and predictable patterns, the difficulty signal offers little advantage over visual features alone.

One factor influencing gaze prediction performance is the use of SMOTE-generated synthetic hard states. While these samples enhance difficulty recognition, they may not fully capture the characteristics of real gameplay frames, leading to a slight reduction in gaze accuracy. Refining synthetic data generation to preserve gaze accuracy while retaining gains in difficulty classification is a promising avenue for improvement.

Future work should also explore applying TAGIL to real-world domains and testing its generalisability across different environments. The current approach benefits from Atari-HEAD's semi-frame-by-frame data collection, where humans had unlimited time to act. In naturalistic settings, such as autonomous driving, rapid reactions can indicate either trivially easy situations or extremely difficult, urgent ones; disambiguating these cases will be essential for deployment beyond controlled environments. In addition, further evaluation across a wider range of game genres and interactive contexts would help assess whether TAGIL's advantages are indeed strongest in maze-like, decision-dense environments and better define its limitations.

Overall, our findings show that incorporating temporal attention and state difficulty into imitation learning can produce agents that more closely replicate human decision-making in complex, dynamic environments. Continued research in this direction has the potential to improve the adaptability and performance of AI systems that learn from human demonstrations.

References

Bain M, Sammut C (1995) A framework for behavioural cloning. In: Machine Intelligence 15, pp 103–129

Banayeeanzade A, Bahrani F, Zhou Y, et al (2025) Gabril: Gaze-based regularization for mitigating causal confusion in imitation learning. arXiv preprint arXiv:250719647

Chawla NV, Bowyer KW, Hall LO, et al (2002) Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16:321–357

Chen R, Kunde GJ, Tao L, et al (2025) Foveal vision reduces neural resources in agent-based game learning. Frontiers in Neuroscience 19:1547264

Farnebäck G (2003) Two-frame motion estimation based on polynomial expansion. In: Image Analysis, SCIA 2003, Proceedings, Lecture Notes in Computer Science, vol 2749. Springer, p 363–370

Ishida Y, Matsubara T, Kanai T, et al (2025) Where do we look when we teach? analyzing human gaze behavior across demonstration devices in robot imitation learning. arXiv preprint arXiv:250605808

Itti L, Koch C, 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 LP, Littman ML, Moore AW (1996) Reinforcement learning: A survey. Journal of artificial intelligence research 4:237–285

Kim H, Ohmura Y, Kuniyoshi Y (2021) Gaze-based dual resolution deep imitation learning for high-precision dexterous robot manipulation. IEEE Robotics and Automation Letters 6(2):1630–1637

Kim H, Ohmura Y, Kuniyoshi Y (2022) Memory-based gaze prediction in deep imitation learning for robot manipulation. In: 2022 International Conference on Robotics and Automation (ICRA), IEEE, pp 2427–2433

Klink P, D'Eramo C, Peters JR, et al (2020) Self-paced deep reinforcement learning. In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020), pp 9216–9227

Lin TY, Goyal P, Girshick R, et al (2017) Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp 2999–3007, https://doi.org/10.1109/ICCV.2017.324

Liu M, Zhao H, Yang Z, et al (2021) Curriculum offline imitation learning. In: Advances in Neural Information Processing Systems 34 (NeurIPS 2021), pp 6266–6277

Mandlekar A, Xu D, Wong J, et al (2022) What matters in learning from offline human demonstrations for robot manipulation. In: Proceedings of the 5th Conference on Robot Learning, Proceedings of Machine Learning Research, vol 164. PMLR, pp 1775–1798

Mnih V, Kavukcuoglu K, Silver D, et al (2015) Human-level control through deep reinforcement learning. nature 518(7540):529–533

Moore AW, Atkeson CG (1993) Prioritized sweeping: Reinforcement learning with less data and less time. Machine learning 13:103–130

Osa T, Pajarinen J, Neumann G, 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, 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 RJ, 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, IEEE, pp 1–8

Saran A, Zhang R, Short ES, et al (2021) Efficiently guiding imitation learning agents with human gaze. In: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, pp 1109–1117

Schaul T (2015) Prioritized experience replay. arXiv preprint arXiv:151105952

Thakur RK, Sunbeam M, Goecks VG, et al (2021) Imitation learning with human eye gaze via multi-objective prediction. arXiv preprint arXiv:210213008

Thammineni C, Manjunatha H, Esfahani ET (2020) Selective eye-gaze augmentation to enhance imitation learning in Atari games. arXiv preprint arXiv:201203145

Torabi F, Warnell G, Stone P (2018) Behavioral cloning from observation. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pp 4950–4957

Wang Y, Xu C, Du B, et al (2021) Learning to weight imperfect demonstrations. In: International Conference on Machine Learning, PMLR, pp 10961–10970

Zhang R, Liu Z, Zhang L, et al (2018) AGIL: Learning attention from human for visuomotor tasks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 663–679

Zhang R, Walshe C, Liu Z, et al (2020) Atari-HEAD: Atari human eye-tracking and demonstration dataset. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 6811–6820