Winning Solution of the BIB MVCS Challenge 2022

Alice Hein and Klaus Diepold

TUM School of Computation, Information and Technology
Technical University of Munich, Germany
alice.hein@tum.de

Abstract. Although AI has made great strides in recent years, state-of-the-art models still largely lack core components of social cognition which emerge early in infant development. The Baby Intuitions Benchmark was designed to compare these "commonsense psychology" abilities in humans and machines. RNN-based models previously applied to this dataset have been shown to not capture the desired knowledge. Here, we apply a different class of deep learning-based models, namely a video transformer. We show that this model quantitatively more closely matches infant intuitions, in that it tends to expect agents' actions to be goal-directed and rational. However, qualitative error analyses show that the model fails to learn the intended causal mechanism underlying instrumental actions, leaving parts of the benchmark an open challenge.

1 Introduction

The foundations of "commonsense psychology" emerge early on in a human's development: Even pre-verbal infants have expectations about agents' goals, preferences and actions [13]. Although deep learning (DL) has made much progress in recent years, this core component of human cognition is still lacking in many state-of-the-art DL models [11]. When tested on the Baby Intuitions Benchmark (BIB), a dataset designed to compare the social cognitive abilities of infants and machines, behavioral cloning (BC) and video prediction models based on recurrent neural networks (RNNs) failed to show infant-like reasoning [5].

Here, we evaluate a different class of DL model, namely a video transformer (VT), on BIB. Recent years have seen the rise of transformers in various areas of AI, including tasks adjacent to social cognition, such as trajectory prediction for cars or pedestrians [16,12,2,14,8,15] and spatial goal navigation [3,1,4]. However, transformer-based video prediction models require many costly pairwise computations. They are usually trained and evaluated on datasets such as *Kinetics-400* or *UCF101*, where video clip lengths range from 7 to 10 seconds – much shorter than those used in BIB, which can be up to 2 minutes long. We therefore implement some modifications to allow a VT to process BIB episodes, and evaluate the resulting model. We find that the VT quantitatively matches infant intuition more closely than previously tested DL baselines. However, qualitative error analyses show that the model does not generalize in the desired way on some of the tasks.

2 Baby Intuitions Benchmark

BIB is a dataset designed to test whether machine learning systems can discern the goals, preferences, and actions of others [5]. It consists of videos in the style of Heider and Simmel's animations [9], where agents, represented by simple shapes, carry out actions in a 2D grid world. BIB follows the violation-of-expectation (VoE) paradigm, i.e., each video has a familiarization and a test phase. The familiarization phase consists of eight successive trials during which an agent consistently displays a certain behavior, allowing the observer to form an expectation of future actions. The test phase includes an expected outcome (perceptually similar to the previous trials, but involving a violation of expectation), and an unexpected outcome (perceptually less similar, but conceptually more plausible).

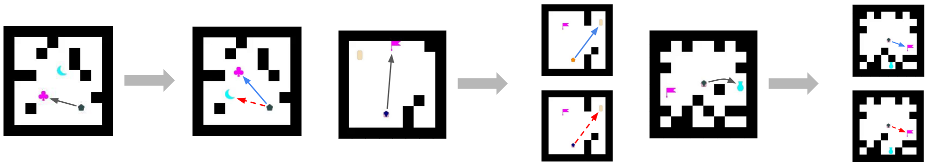
Because BIB adopts its tasks and paradigm from developmental cognitive science and provides sufficient data to train DL-based models, it allows for the direct comparison of human and machine performance [5]. A critical first step in this direction was taken by Stojnic et al. [13], who collected infant responses on a representative selection of BIB episodes and compared them with three state-of-the-art DL models from two classes: Behavioral cloning (BC) and video modeling. Recently, Zhi-Xuan et al. [17] proposed a principled alternative to DL approaches, based on a hierarchically Bayesian Theory of Mind (HBToM). Results from both works serve as comparisons in this paper.

2.1 BIB Tasks

Goal-directed actions The *preference* task (1,000 episodes) tests whether an observer represents agents as having a preference for goal objects, rather than locations. The setup consists of two goals and an agent, whose starting position is fixed. In the familiarization trials, the agent consistently moves towards the same object. Goal locations and identities are correlated, such that preferred and nonpreferred goals have a similar position across trials. In the test phase, the two objects appear in positions previously seen during familiarization. However, goal identities are switched. In the expected outcome, the agent moves to the preferred object. In the unexpected outcome, the agent follows the same trajectory as seen during familiarization and moves to the nonpreferred object (see Figure 1a).

The *multi-agent* task (1,000 episodes) tests whether an observer attributes specific goal preferences to specific agents. The setup consists of two goal objects appearing at different positions across trials and an agent with a fixed starting position. Again, the agent moves repeatedly to the same object during familiarization. In the unexpected test outcome, the agent moves towards its nonpreferred goal. In the expected outcome, a new agent replaces the previously seen one and moves toward the familiar agent's nonpreferred object. The unfamiliar agent choosing a new goal should be less surprising than the familiar agent switching preference (see Figure 1b).

The *inaccessible-goal* task (1,000 episodes) tests whether an observer understands the principle of solidity, and that physical obstacles may restrict agents' actions. The familiarization trials are identical to the *multi-agent* task. In the expected test trial, the previously preferred object is made inaccessible by a barrier, and the agent moves to the other goal. In the unexpected test trial, the agent switches preference despite both objects remaining accessible (see Figure 1c).



(a) Example of a *preference* task. Goal locations are switched for testing. In the expected outcome, the agent still chooses the same object. In the unexpected outcome, the agent instead follows the familiar path to its nonpreferred goal.

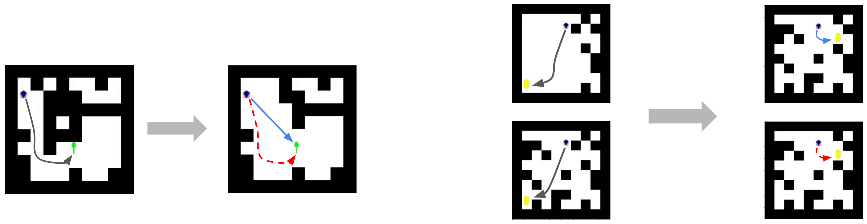
(b) Example of a *multi-agent* task. A new agent appears in the test trial. This new agent choosing the other agent's nonpreferred object (top right) should be less surprising than the familiar agent doing so (bottom right).

(c) Example of an *inaccessible-goal* task. The agent switches goals in the test trial. This should be expected if the preferred object is inaccessible (top right), but unexpected if both objects are accessible (bottom right).

Fig. 1: Examples of goal-directed action tasks. Agents move repeatedly to the same goal during familiarization (left), while test trials differ by task type (right). Blue solid lines represent expected outcomes, red dashed lines represent unexpected outcomes.

Efficient actions The *efficient-agent* task tests whether an observer expects agents to move efficiently towards their goal. It consists of two subtasks: path control (1,500 episodes) and time control (1,000 episodes). In both subtasks, the setup consists of one goal object and one agent. During familiarization, the agent moves efficiently towards the object, but must navigate around a barrier to reach it. This obstacle is removed in the test phase. In both subtasks, the expected outcome consists of the agent moving efficiently towards its now-unobstructed goal. For the path control task, a previously seen combination of agent and goal location is used, and the unexpected outcome consists of the agent moving along the familiar, but now inefficient, trajectory (see Figure 2a). For the time control subtask, the goal object is placed closer to the agent and the unexpected outcome consists of the agent following a path that is inefficient, but takes up the same amount of time as the efficient one.

The *inefficient-agent* task (890 episodes) tests whether an observer forms expectations about the actions of irrational agents. During familiarization, an agent is shown either moving efficiently, as in the *efficient-agent* task, or inefficiently. In the test phase, the agent is shown moving inefficiently to the goal. This should be an unexpected outcome if the agent previously behaved rationally and an expected outcome if the agent previously behaved irrationally (see Figure 2b).



(a) Example of an *efficient-agent* task. During familiarization, the agent navigates efficiently around an obstacle to reach its goal. The barrier is removed during testing. The agent is expected to now move efficiently, rather than following the same path as before.

(b) Example of an *inefficient-agent* task. An agent that moves inefficiently during familiarization (top) is expected to continue doing so during testing, whereas an efficient agent (bottom) beginning to move inefficiently should be surprising.

Fig. 2: Examples of efficiency tasks. Familiarization trial shown on the left, test trials on the right. Blue solid lines represent expected outcomes, red dashed lines represent unexpected outcomes.

Instrumental actions The *instrumental-action* task (987 episodes) tests whether an observer can recognize an agent's action sequences as instrumental and directed towards higher-order goals. The setup consists of a goal, an agent, a removable green barrier with a lock, and a key, represented by a red triangle. During familiarization, the goal is obstructed by the green barrier. The agent collects the key, inserts it into the lock, removes the barrier, and moves to the goal. In the test phase, a key is still present, but the green barrier is either absent or no longer blocking the goal. In the expected outcome, the agent moves directly towards the goal, whereas it still moves towards the now-obsolete key in the unexpected outcome (see Figure 3).

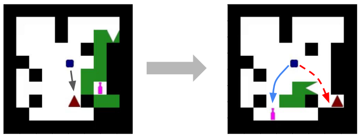


Fig. 3: Example of an *instrumental-action* task.

Background training episodes To facilitate the training of machine learning models, BIB includes a large number of background episodes which share the same structure, agents, and goal objects as the test set. However, only expected trials are provided during training. The training set is divided into four tasks. To generalize systematically to the test trials, the model needs to combine knowledge acquired from all four training tasks. In the *single-object* task (10,000 episodes), an agent navigates efficiently to a goal object (see Figure 4a). In the *preference* task (10,000 episodes), the agent consistently chooses one object over another in all trials (see Figure 4b). In contrast to the *preference* test task, both objects are very close to the agent, so navigation is not trained. In the *multi-agent* task (4,000 episodes), the agent moves to a single object of very close proximity (see Figure 4c). At some point during the episode, the agent is replaced with a new agent. This differs from the *multi-agent* test task, where there are two goals which are placed farther away and the new agent only appears in the test trial. In the *instrumental-action* task (4,000 episodes), the agent is initially confined by a green barrier, which it removes with a key in order to move to its goal. This differs from the *instrumental-action* test task in that the barrier surrounds the agent, rather than the goal.

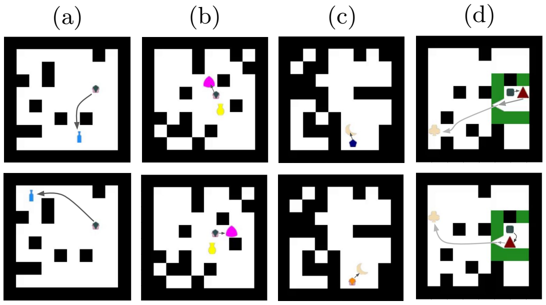


Fig. 4: Examples of training trials consisting of *single-object* (4a), *preference* (4b), *multi-agent* (4c), and *instrumental* (4d) tasks.

3 Model

3.1 Architecture

Our model consists of a convolutional neural network (CNN) encoder, a transformer component, a CNN decoder, and a feedforward output layer. A schematic visualization is shown in Figure 5. The CNN encoder has two convolutional layers and two max-pooling layers. For each $3 \times 84 \times 84$ input image, it produces a $30 \times 21 \times 21$ representation, which we concatenate with x- and y-position encodings, resulting in $32 \times 21 \times 21$ image patches.

The transformer component consists of three standard five-layer attention blocks with 8 heads of input dimension 32 and hidden dimension 256. The first block performs cross-attention over the test trial's encoded first frame and the previous familiarization trials. Because attending over every patch, frame, and trial would require more memory than was available to us, we only feed in the top k patches per frame that display the highest change compared to the previous frame. K was set to 3, because this provided a satisfactory balance between computational complexity and performance. The results of attending over each trial are then averaged and passed through a self-attention block, followed by another cross-attention block. This cross-attention block attends over past steps in the test trial, encoded in the same way as the familiarization trial frames.

In a final step, the outputs of the transformer component are passed through a linear layer, which produces a $1 \times 21 \times 21$ prediction of the agent's next position, and a CNN decoder, which produces a $3 \times 83 \times 84$ prediction of the video's next frame.

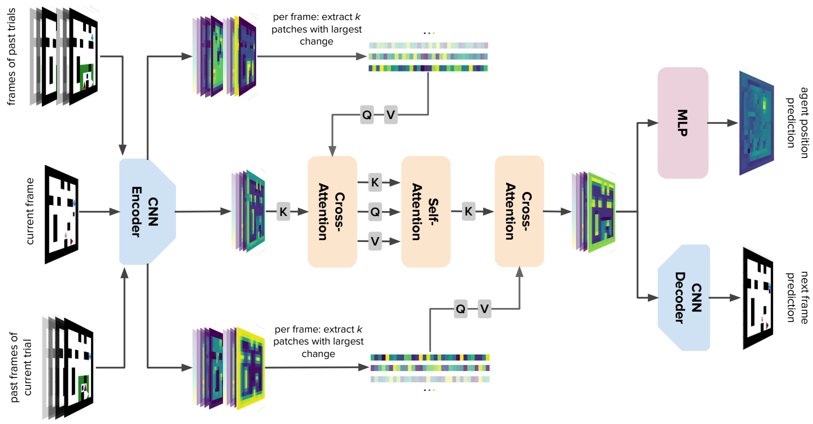

frames of past trials
current frame
past frames of current trial
CNN Encoder
per frame: extract k patches with largest change
per frame: extract k patches with largest change
...
...
Q
V
K
Cross-Attention
K
Q
V
Self-Attention
K
Cross-Attention
Q
V
MLP
agent position prediction
CNN Decoder
next frame prediction

Fig. 5: Schematic visualization of the VT architecture.

3.2 Training and Testing

As in Gandhi et al. [5], the videos' frame rate was downsampled by a factor of 5. We used a maximum sequence length of 90. Frame rates of longer sequences were interpolated to fit the maximum length. Of the background episodes of BIB, we used 80% for training, 15% for testing, and 5% for validation. Models were trained using the Adamax optimizer for a total of 6 epochs. The batch size was set to 6 because of the VT's high memory requirements. We tested the models on the validation set in five evenly spaced intervals per epoch and saved the model with the lowest validation loss to avoid overfitting.

Our loss function consisted of the sum of two terms. The first term was the binary cross-entropy (BCE) loss between the prediction of the agent's next step and the actual agent position. To address the imbalance between the "agent" and "no-agent" class, we employed a weighted version of the BCE loss, which is widely used in instance segmentation [10]. The second term was the mean squared error (MSE) between the prediction of the next frame and the actual next frame, upweighted by a constant factor so that both loss terms were scaled evenly. This second term was introduced because transformers may disregard agent identities unless incentivized otherwise [16]. For tasks like *preference*, which is based on the preservation of agent shapes and colors, we, therefore, found that it improved performance to include an auxiliary reconstruction loss. During evaluation, only the main BCE loss was used.

On a 16-Core AMD EPYC 7282 server with six GeForce RTX 2080 GPUs, the training time was around 3 hours per epoch. Our code is available at `https://github.com/zero-k1/BIB-VT`.

4 Results

In total, we trained five models with different random seeds, and we report their average performance and standard deviations. The baseline DL models previously tested on BIB used the prediction error of the frame with the highest loss as their metric of “surprise”, as this provided better results compared to the mean error over entire trials [5]. In our case, the mean error yielded a higher performance on most tasks, which is why we report both metrics here. Performance comparisons with models previously tested on BIB are shown in Table 1. However, binary VoE accuracies include no information about the magnitude of the difference in surprisal scores between expected and unexpected trials. We therefore also show z-scored means of both the models’ average prediction error and infants’ looking times, as reported by Stojnic et al. [13], in Figure 6.

Table 1: VoE Accuracy on BIB tasks. VT (Mean) uses the avg. error over all test trial frames as the ”surprise” metric, whereas VT (Max) uses the error for the frame with the highest loss. Baselines and Video Transformers are deep learning-based, whereas HBToM uses a principled Bayesian solution.

		Baselines			Video Transformer (ours)	
Task	**HBToM**	**BC-MLP**	**BC-RNN**	**Video-RNN**	**VT (Mean)**	**VT (Max)**
Goal-directed						
Preference	99.7	26.3	48.3	47.6	82.1 ± 0.0	80.8 ± 0.0
Multi-Agent	99.2	48.7	48.2	50.3	49.1 ± 0.0	49.2 ± 0.0
Inaccessible	99.7	76.9	81.6	74.0	89.8 ± 0.0	85.5 ± 0.0
Efficiency						
Path Control	94.9	94.0	92.8	99.2	97.3 ± 0.0	97.5 ± 0.0
Time Control	97.2	99.1	99.1	99.9	99.8 ± 0.0	99.7 ± 0.0
Irrational	96.6	73.8	56.5	50.1	29.5 ± 0.1	34.1 ± 0.1
Instrumental Actions						
No Barrier	98.8	98.8	98.8	99.7	98.7 ± 0.0	97.9 ± 0.0
Inconsequential Barrier	97.0	55.2	78.2	77.0	96.9 ± 0.0	91.9 ± 0.0
Blocking Barrier	99.7	47.1	56.8	62.9	82.1 ± 0.1	64.2 ± 0.1

4.1 Goal-directed

Preference In contrast to previous DL-based models, the VT seems, at least to some degree, to associate agents with certain goal preferences in the *preference* task (see Figure 6a). To investigate which parts of the familiarization trials the model relied most on for its decisions, we performed an occlusion analysis. We used only one trial as the familiarization input (performance was almost identical when using one vs. the full eight trials), and dropped each of the patches fed into the first transformer block in turn. For each patch, we recorded the z-scored difference in prediction error between the expected and unexpected outcome. An example result is shown in Figure 7. Models tended to either focus on the agent’s last or first step. Averaged over all models and episodes, the patch with the largest impact on the final prediction was part of the last two frames of the familiarization trial in 52.6% of cases.

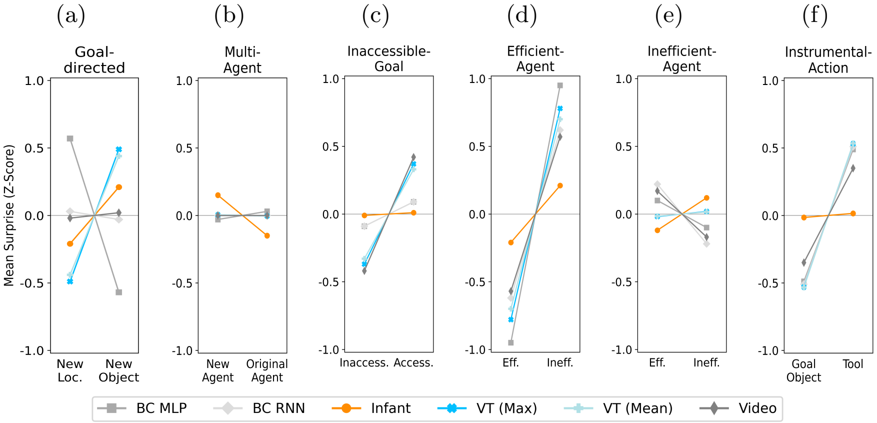


Fig. 6: Z-scored means of the average surprisal scores of the models and the looking times of the infants to the expected and unexpected outcomes in the BIB test episodes.

Multi-Agent Similar to the other DL models, the VT does not acquire the desired knowledge from the *multi-agent* background training tasks, which feature both agents moving towards the same single goal across trials. Note that infants tested on BIB were, in fact, more surprised at the supposedly ”expected” trials (see Figure 6b). Stojnic et al. [13] hypothesize that this may be due to the increased novelty of the new agent. A closer look at the frame predictions produced by the VT hints at some confusion regarding the agents' identity: In some cases, the model reconstructs the familiar agent in the unexpected trial, rather than the new agent present in the input (see Figure 8 for an example). Averaged over all models and episodes, this was the case 27.9% of the time.

Inaccessible In the *inaccessible-goal* task, the VT model achieves a higher accuracy than previous DL models. It exhibits a stronger deviation in surprise than the infants, who were indifferent on this task (see Figure 6c). Stojnic et al. [13] posit that infants may have considered the new barrier in the expected outcome as indicative of a new environment and did not carry over any expectations of goal preference from the familiarization trials. Although the VT has a lower prediction loss on the expected outcome in most cases, it is more ”split” than in the single-object case (see Figure 9 for an example prediction). Averaged over all models and episodes, the entropy of the models' prediction on the test trial's last frame was 1.10 for the expected, and 1.47 for the unexpected outcome. For comparison, the average entropy for the last frame of the single-object *efficiency-time* trials was only 0.58.

4.2 Efficiency

Similar to previous models, the VT's VoE accuracy on the *path-control* and *time-control* tasks is nearly perfect – the model strongly expects agents to move

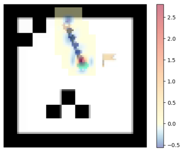


Fig. 7: Z-scored impact of omitting a patch from the familiarization trial.

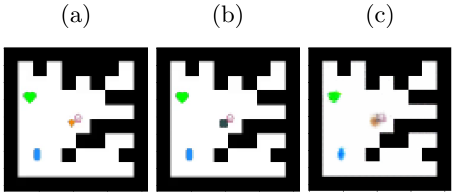


Fig. 8: (a) Unexpected *multi-agent* outcome (familiar agent). (b) Expected outcome (new agent). (c) Prediction for expected outcome.

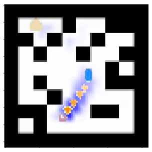


Fig. 9: *Inaccessible goal* task. Predicted agent positions marked blue.

towards their goal efficiently. This is in accordance with infants' intuitions (see Figure 6d). On the *inefficient-agent* task, the VT tends to be more surprised at the previously inefficient model moving inefficiently than at the previously efficient agent doing so. Although not necessarily a desired outcome, this is actually more in line with the intuitions of the infants tested on BIB, who attributed rational action both to previously efficient and inefficient agents in a new environment (see Figure 6e). When we compare the impact of the familiarization trials featuring the efficient vs. inefficient agent on the VT model (see Figure 10), we see that a similar mechanism is at work: The lowest levels, which attend over past familiarization trials, show differences in activation. However, these differences disappear almost completely throughout the higher layers. This leads to the inefficient agent being treated in the same way as the efficient one, which explains that the mean surprise score is almost the same in both cases. The slightly larger error for the inefficient agent probably stems from the fact that irrational agents are not seen during training, leading to higher prediction uncertainty.

4.3 Instrumental Actions

Compared to the other DL models, the VT performs similarly on episodes without barriers and better on episodes with inconsequential or blocking barriers. Again, infants were indifferent on this task (see Figure 6f). Stojnic et al. [13] note that they may have failed to recognize the instrumental actions because they were causally opaque. Although the VT is correct in most cases in terms of VoE accuracy, it also does not seem to fully understand the causal mechanism. A look at the frame predictions shows that the model usually expects the disappearance of the key on the first step, even though the agent has not collected and inserted it. Averaged over all models and episodes, the VT at least partly predicts the key's position as the agent's first step in 47% of cases, even though the key is mostly far away from the agent. This is most likely because the key is always right next to the agent in the background *instrumental-action* tasks, and thus constitutes its first step. The VT also often predicts the disappearance of the green barrier towards the end of the episode, even though the key was not inserted. This is most likely because the green barrier has always disappeared by the time the agent reaches the goal in the background tasks. Occlusion analyses support this hypothesis: The parts of the test trial that contribute the most to the z-scored MSE prediction error on expected *instrumental-action* outcomes were usually the first and last steps of the agent (see Figure 11 for an example).

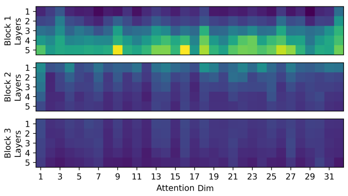


Fig. 10: Avg. difference in the VT layers' activations when processing the episodes' unexpected vs. expected familiarization trials, featuring a rational or an irrational agent, respectively.

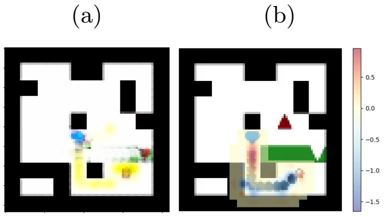


Fig. 11: Prediction on an *instrumental-action* task. (a) Predicted last frame and agent trajectory (yellow). (b) Z-scored impact of each test trial patch on final MSE error.

5 Discussion and Conclusion

In conclusion, the VT model tested in this paper outperforms previous baselines based on DL in the *preference*, *inaccessible-goal*, and *instrumental-actions* BIB tasks in terms of VoE accuracy. Its surprisal scores are also more in line with infants' expectations than previous DL models, in that it tends to represent agents' actions as directed towards goals, rather than locations, and defaults to expecting rational actions. This suggests that the transformer's attention mechanism can be helpful in acquiring intuitions about agents' goals, preferences, and actions purely from predicting the next step in videos.

However, a qualitative analysis of the VT's errors also demonstrated the pitfalls of this approach: Models may exploit the particularities of a training dataset in an unintended way [6,7], e.g. by associating the disappearance of the green barrier in the *instrumental-actions* task with the agent's first and last step rather than the key mechanism, thus failing to generalize on out-of-distribution data. This may be mitigated with a more realistic data setting, where models can gain experience with diverse agents and interactively disambiguate the causes and effects of instrumental mechanisms in a manner closer to human infants. The findings also support the benefit of investigating hybrid architectures that incorporate methods that explicitly model human intuitions, such as HBToM, to take advantage of both the flexibility of DL-based approaches and the data efficiency and robustness of principled Bayesian models.

References

1. Chaplot, D.S., Pathak, D., Malik, J.: Differentiable spatial planning using transformers. In: International Conference on Machine Learning. pp. 1484–1495. PMLR (2021)
2. Chen, W., Wang, F., Sun, H.: S2tnet: Spatio-temporal transformer networks for trajectory prediction in autonomous driving. In: Asian Conference on Machine Learning. pp. 454–469. PMLR (2021)
3. Du, H., Yu, X., Zheng, L.: Vtnet: Visual transformer network for object goal navigation. arXiv preprint arXiv:2105.09447 (2021)
4. Fukushima, R., Ota, K., Kanezaki, A., Sasaki, Y., Yoshiyasu, Y.: Object memory transformer for object goal navigation. arXiv preprint arXiv:2203.14708 (2022)
5. Gandhi, K., Stojnic, G., Lake, B.M., Dillon, M.R.: Baby Intuitions Benchmark (BIB): Discerning the goals, preferences, and actions of others. Advances in Neural Information Processing Systems **34**, 9963–9976 (2021)
6. Gardner, M., Artzi, Y., Basmov, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., et al.: Evaluating models' local decision boundaries via contrast sets. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1307–1323 (2020)
7. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
8. Giuliari, F., Hasan, I., Cristani, M., Galasso, F.: Transformer networks for trajectory forecasting. In: 2020 25th international conference on pattern recognition (ICPR). pp. 10335–10342. IEEE (2021)
9. Heider, F., Simmel, M.: An experimental study of apparent behavior. The American journal of psychology **57**(2), 243–259 (1944)
10. Jadon, S.: A survey of loss functions for semantic segmentation. In: 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). pp. 1–7. IEEE (2020)
11. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and brain sciences **40** (2017)
12. Li, L.L., Yang, B., Liang, M., Zeng, W., Ren, M., Segal, S., Urtasun, R.: End-to-end contextual perception and prediction with interaction transformer. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5784–5791. IEEE (2020)
13. Stojnic, G., Gandhi, K., Yasuda, S., Lake, B.M., Dillon, M.R.: Commonsense Psychology in Human Infants and Machines (Jun 2022). https://doi.org/10.31234/osf.io/j3zs8, psyarxiv.com/j3zs8
14. Sui, Z., Zhou, Y., Zhao, X., Chen, A., Ni, Y.: Joint intention and trajectory prediction based on transformer. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7082–7088. IEEE (2021)
15. Yu, C., Ma, X., Ren, J., Zhao, H., Yi, S.: Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In: European Conference on Computer Vision. pp. 507–523. Springer (2020)
16. Yuan, Y., Weng, X., Ou, Y., Kitani, K.M.: Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9813–9823 (2021)
17. Zhi-Xuan, T., Gothoskar, N., Pollok, F., Gutfreund, D., Tenenbaum, J.B., Mansinghka, V.K.: Solving the baby intuitions benchmark with a hierarchically bayesian theory of mind. arXiv preprint arXiv:2208.02914 (2022)