MAFEA: Multimodal Attribution Framework for Embodied AI

Anonymous Author(s)
Affiliation
Address
email

Abstract

Understanding multimodal perception for embodied AI is an open question because such inputs may contain highly complementary as well as redundant information for the task. A relevant direction for multimodal policies is understanding the global trends of each modality at the fusion layer. To this end, we disentangle the attributions for visual, language, and previous action inputs across different policies trained on the ALFRED dataset. Attribution analysis can be utilized to rank and group the failure scenarios, investigate modeling and dataset biases, and critically analyze multimodal EAI policies for robustness and user trust before deployment. We present MAFEA, a framework to compute global attributions per modality of any differentiable policy. In addition, we show how attributions enable lower-level behavior analysis in EAI policies through two example case studies on language and visual attributions.

Dual submission: NeurIPS 2022, 5th Robot Learning: Trustworthy Robotics.

1 Introduction

Embodied AI policies have achieved remarkable success in simulated 3d environments [1, 2, 3, 4] and physical robots [5, 6]. Analogous to the success of end-to-end learning for image classification [7] and language modeling [8], recent works on end-to-end embodied policies [9] attempt to solve everyday complex tasks, like 'making a cup of coffee' or 'throwing a chilled tomato slice in trash' (as shown in Fig. 1). Such policies often fail for reasons that are poorly understood. Interpreting the decision-making in end-to-end policies is important to enable trustworthy deployment and handle failure case scenarios.

Learning an embodied AI policy typically involves function f mapping the observation/state O_t to an action a_t. In many environments [10, 11], the observation can just be the visual frame at current timestep $O_t = \{V_t\}$. However, complex tasks can hardly be solved with just a single type of observation. In Atari [12], the policy operates on last 4 visual frames: $O_t =$

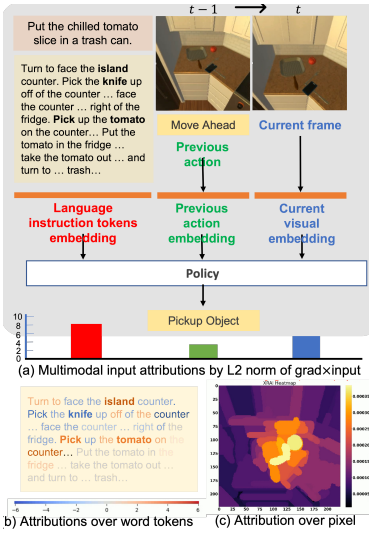


Figure 1: Attribution analysis of multimodal EAI policy. (a) (left to right) shows attribution for language, previous action and visual frame. (b) **Words**: red is positive attribution, blue is negative, grey is neutral; (c) **Pixels**: yellow is high attribution, purple is low.

Submitted to NeurIPS 2022 Progress and Challenges in Building Trustworthy Embodied AI Workshop (TEA 2022). Do not distribute.

$\{V_{t-3}, V_{t-2}, V_{t-1}, V_t\}$. For long horizon tasks in a large sparse maze environment [13], inputs may include the previous action: $O_t = \{V_t, a_{t-1}\}$, as it will be likely that the next action is to 'move forward' if the previous action was the same. Real-world robotics navigation relies on visual frame and proprioception [6] to successfully navigate undetected obstacles. In this work, we consider a mobile robot that takes natural language instructions and interacts with objects in simulated household environments [9]. Such a robot needs to predict sequence of actions that would complete the task, given the previous action a_{t-1}, natural language instructions L and ego-centric vision V_t, that is the observation $O_t = \{V_t, a_{t-1}, L\}$. While many attempts have been made to train policies for this task [9, 14, 15, 16], the existing best performance is far below that of an average human [9].

To this end, we investigate how multimodal a policy is, in terms of attribution given to visual, language, and previous actions. Aggregated analysis per modality provides insights on the effect of modeling choices, such as at what layer the fusion happens, what sub-network is used to process each modality separately, and how the representation per modality affects the fusion process. Further, attributions provide an introspection technique to rank the contributions from each modality in a decision, as well as analyze modeling and dataset biases. Our main contribution is to propose MAFEA for (a) global analysis in terms of the percentage of average attribution per modality, as well as (b) modality-specific local case studies.

2 Related Work

Interpretability and explainability Recent work in multimodal explainability in autonomous vehicles [17] uses symbolic explanations to debug and process outputs out of sub-components. In contrast, we address the challenge of post-hoc multimodal interpretability for any existing end-to-end trained differentiable policies. GRAD $\odot$ INPUT [18], a simple and modality-agnostic attribution that works on par with recent methods [19]. We use this method to compute multimodal attribution at inputs to the fusion layer to weigh how each modality contributes to the decision-making. While GRAD $\odot$ INPUT is a modality-agnostic starting point for attributions, it is not easy to understand, especially for images. Among recent works to improve visual attribution [20, 21, 22, 23, 24, 25, 26], we use XRAI [26] for vision-specific analysis as it produces visually intuitive attributions by relying on regions, not individual pixels.

Language-driven task benchmarks There are many benchmarks to study an agent's ability to follow natural language instructions [9, 27, 28, 29]. ALFRED [9] serves as a suitable testbed for this analysis as these tasks require both high reasoning for navigation and manipulation. ALFRED dataset provides visual demonstrations collected through PDDL planning in 3D Unity household environments and natural language description of the high-level goal and low-level instructions annotated by MTurkers. The benchmarks provide evaluation metrics for the overall task goal completion success rate (SR) and those weighted by the expert's path length (PLWSR) and have reported a huge gap in the performance of learning algorithms and humans at these tasks.

End-to-end Learned Policies We investigate the end-to-end learned policies for the task, such that, the gradient can be attributed at a task level. While we do not discuss modular yet differentiable policies like [16] [30], tying the gradient across multiple modular learned components is a direction for future work. In our work, we consider the checkpoints of policies trained on the ALFRED dataset. Broadly, these policies are of two types: (a) sequence-to-sequence models, that are, the one proposed with ALFRED dataset (Baseline) [9] and Modular Object-Centric Approach (MOCA) [14], (b) transformer-based models, that are Episodic Transformers (ET) [15], and Hierarchical Tasks via Unified Transformers (HiTUT) [31]. Refer Table 1 to compare architectural details [1].

3 Approach

Consider a linear model $f(x) = w_1x_1 + w_2x_2$. To explain which feature is more important for predicting the value of $f(x)$, we can compare their coefficients. If $w_1 = 1000$ and $w_2 = 0.01$, we can say that x_1 would be weighed more than x_2. Gradients are the general way of discussing the coefficient of the importance of a particular feature in deciding the output. Using only weights or gradients as an attribution assumes that the values of x_1 and x_2 are of the same order, like image

[1] Previous action is modeled with learned embedding look-up in all these policies.

Table 1: Policies trained on ALFRED Dataset and their architectures for each modilty

Policies	Visual	Language	Fusion
Baseline [9]	Frozen ResNet-18	Learned Embedding, Bi-LSTM	LSTM
MOCA [14]	Frozen ResNet-18 + Dynamic Filters	Learned Embedding, Bi-LSTM	LSTM with residual connection
ET [15]	Frozen ResNet-50	Learned Embedding, Transformer	Transformer Encoder
HiTUT [31]	Frozen MaskRCNN	Learned Embedding, FC, LN	Transformer Encoder

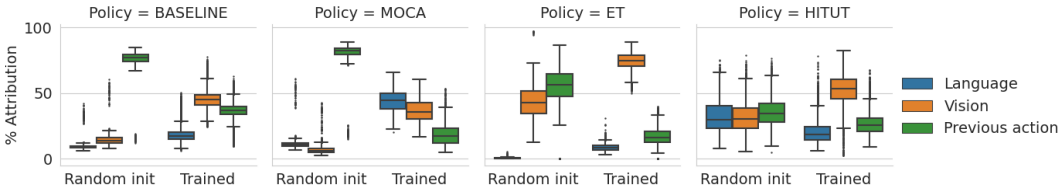


Figure 2: % Multimodal Attribution for policies *randomly initialized* and *trained* on ALFRED dataset. The performance of the policies improves from left to right. Trained checkpoints are evaluated with 100 trajectories sampled from validation seen trajectories. Policies are randomly initialized for 5 seeds each and evaluated over 10 validation seen trajectories.

pixels. However, this doesn't hold in the case of multimodal policies when x_1 is an image embedding and x_2 is a language embedding. *Element-wise product of gradient into input*, also known as GRAD $\odot$ INPUT [18], provides global importance about the input feature in the dimensionality of the input feature itself. We compute the attributions per modality with respect to the predicted action label output. The differentiable policy $f : (V_t, a_{t-1}, L) \rightarrow a_t$ takes input as current visual frame V_t, previous action label embedding a_{t-1}, and the language instruction L. At the penultimate fusion layer, an intermediate representation is typically a vector or matrix. The attribution for each modality can be computed by GRAD $\odot$ INPUT, where input is the feed-forward features computed at the penultimate fusion layer. Let the policy neural network be f that outputs a softmax distribution over the action to be taken. The input feature attribution α_i for i^{th} dimension of vector $\mathbf{x}$ for the most likely predicted action is computed as $\alpha_i = \frac{\partial(\max f(x))}{\partial x_i} \odot x_i$. For each modality represented as a vector, we need to pool the attribution per dimension to compute a scalar value attribution. Here are the implications of different pooling approaches:

L^∞	gives the maximum magnitude value, independent of the dim d (same as L^∞/d).
L^1	provides the sum of all attribution magnitudes.
L^1/d	is same as L^1, but invariant of the dim d.
L^2	diminishes the majority insignificant but non-zero attributions.
L^2/d	has same impact of L^2 but with undesirable scaling by dim d.

To compute modality-specific attribution for latent vectors, L^2 is suitable to include the attributions from every dimension in the vector (unlike L^∞) and reduce the impact of insignificant close to zero attributions (unlike L^1). We do not consider L^1/d or L^2/d as the attribution of modality should depend on the number of latent features allocated to highlight modeling biases.

4 Global Attribution Analysis

To understand how the multimodal attribution at the fusion layer is, we analyze the GRAD $\odot$ INPUT with L2 norm as pooling to compute attributions. In Fig. 2, we show the percentage of attribution given to language, vision, and previous action before and after training by different policies on the ALFRED dataset. To get the attributions for the former, we use randomly initialized policies over 5 seeds. For the trained attributions, we randomly sample 100 trajectories from 820 validation data from seen environments.

Baseline and HiTUT have a balanced attribution after training over all three modalities, with a preference for visual features. While MOCA prefers language slightly over visual and previous action, ET strongly prefers visual features over previous action and language. The attributions prior to

training are used to analyze modeling biases. At initialization, Baseline and MOCA have the majority of the % attribution on the previous action. ET model starts with the majority of the attribution over previous action and visual frame, and very little focus on the language. HiTUT has a balanced attribution before training, which does not inherently induce modality bias.

We compute these attributions based on the expert trajectories to keep the analysis on the more relevant input states within the training distribution. We do not analyze the rollout trajectories because most policies have poor success rates (SR) and are often stuck in irrelevant out-of-distribution states, such as facing a wall. We take the gradients with respect to the most likely action predicted by the policy, and not the expert's chosen action. More information about why we choose the predicted action is described in Appendix B.

5 Modality-specific Attribution Analysis

Does language attribution change depending on the steps taken in the episode? For language attribution, we compute the gradient of the predicted output with respect to embeddings of the word tokens in the language instruction, $W := \texttt{Embedding}(\texttt{Tokenize}(L))$, where $W \in \mathbb{R}^{n \times d}$. We compute the attribution for the matrix W by $\alpha_W := \frac{\partial f}{\partial W} \odot W$. As word embedding is a d-dimensional vector, the gradient is also computed with respect to each feature dimension. To obtain an attribution value for each of the n word tokens in the instruction, we reduce the ith word attribution $\alpha_w = \alpha_W[i]$ of embedding size d, where w is a word token, to a scalar by choosing the value which has the maximum absolute value as the attribution as $\alpha_w := \texttt{max abs}(\alpha_W[i]), \forall i \in n$.

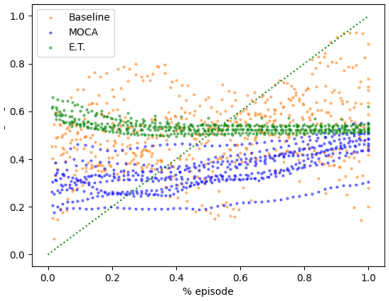


Figure 3: [Global attribution] Focus center of Language attributions per episode length: Baseline [9] (yellow), MOCA [14] (purple) and ET [15] (green) on validation seen dataset.

To see how the focus center of the attribution for language instructions shifts in terms of the percentage of episode completion (Fig. 3[2]), we compute the word-index weighted attribution over all the word tokens in instruction per trajectory $c_{raw} = [(i+1) * \text{abs}(\alpha_w)] \forall i \in n$. To compare across variable length trajectories, we normalize it as $c_{scaled} = \sum c_{raw} / (\sum \text{abs}(\alpha_w) * n)$.

In the ideal case, the focus center of attribution on instruction should increase as the episode nears completion. Fig. 3 shows MOCA (blue), that had high language attribution in Fig. 2, follows a linearly increasing trend. Baseline and ET do not show such a trend, which aligns with Fig. 2 where they show low attribution on language.

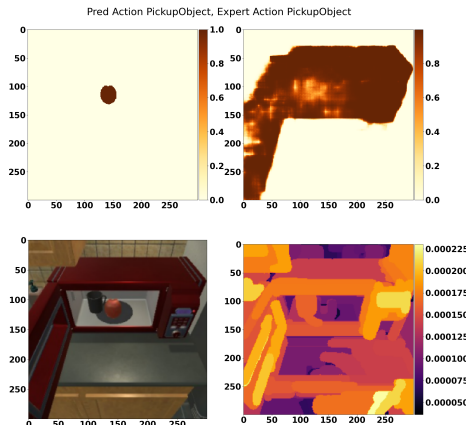


Figure 4: Visual attributions for an interact action can indicate the focus of the baseline policy. Top-Left: Expert's interaction mask, Top-Right: Predicted interaction mask, Bottom-Left: visual observation in RGB, Bottom-Right: XRAI attributions

Does visual attribution align with the predicted interaction mask? We analyze how visual attributions for interact action (PickObject) can indicate the focus of the interaction mask. We calculate the visual attribution with XRAI [26] as shown in Fig. 4 for Baseline policy. While the action label is predicted correctly by the policy, the intersection over union of the mask is small. The simple visualization of the interaction mask reveals that a lot of objects (microwave, mug, and tomato) are selected. With XRAI attribution, we can qualitatively analyze the regions in terms of high (eg. parts of the microwave) and low (eg. tomato) attribution.

[2]HiTUT is not shown in Fig. 3 because it doesn't consider the previous language instructions to make the current prediction which is different from how the other papers do it.

6 Conclusion

In this work, we draw the community's attention to attribution analysis for interpreting multimodal policies. We provide the framework MAFEA for attribution analysis to gain insights into multimodal embodied AI policies. We compare seq2seq (baseline, MOCA) and transformer-based (ET, HiTUT) policies trained on the ALFRED dataset and highlight the modality biases in these models. We also analyze how the focus center in language instructions moves as the episode progresses, and discuss how visual attributions can be used for analyzing successful/unsuccessful action predictions.

References

[1] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *International Conference on Learning Representations (ICLR)*, 2020.

[2] Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *Neural Information Processing Systems*, 2020.

[3] Abhishek Das, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Neural Modular Control for Embodied Question Answering. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2018.

[4] Wei Yang, X. Wang, Ali Farhadi, Abhinav Kumar Gupta, and Roozbeh Mottaghi. Visual semantic navigation using scene priors. In *International Conference on Learning Representations (ICLR)*, 2019.

[5] Mohammadreza Sharif, Deniz Erdogmus, Chris Amato, and Taskin Padir. End-to-end grasping policies for human-in-the-loop robots via deep reinforcement learning*. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2768–2774, 2021.

[6] Zipeng Fu, Ashish Kumar, Ananye Agarwal, Haozhi Qi, Jitendra Malik, and Deepak Pathak. Coupling vision and proprioception for navigation of legged robots. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1894–1904, 2022.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[8] Robert Dale. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118, 2021.

[9] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[10] Charlie Beattie, Joel Z. Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, Julian Schrittwieser, Keith Anderson, Sarah York, Max Cant, Adam Cain, Adrian Bolton, Stephen Gaffney, Helen King, Demis Hassabis, Shane Legg, and Stig Petersen. Deepmind lab. *ArXiv*, abs/1612.03801, 2016.

[11] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

[12] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *ArXiv*, abs/1312.5602, 2013.

[13] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. `https://github.com/maximecb/gym-minigrid`, 2018.

[14] Kunal Pratap Singh, Suvaansh Bhambri, Byeonghwi Kim, Roozbeh Mottaghi, and Jonghyun Choi. Factorizing perception and policy for interactive instruction following. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1868–1877, 2021.

[15] Alexander Pashevich, Cordelia Schmid, and Chen Sun. Episodic Transformer for Vision-and-Language Navigation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[16] So Yeon Min, Devendra Singh Chaplot, Pradeep Kumar Ravikumar, Yonatan Bisk, and Ruslan Salakhutdinov. FILM: Following instructions in language with modular methods. In *International Conference on Learning Representations*, 2022.

[17] Leilani H. Gilpin, Vishnu Penubarthi, and Lalana Kagal. Explaining multimodal errors in autonomous vehicles. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, 2021.

[18] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *ArXiv*, abs/1605.01713, 2016.

[19] Marco Ancona, Enea Ceolini, A. Cengiz Öztireli, and Markus H. Gross. A unified view of gradient-based attribution methods for deep neural networks. *ArXiv*, abs/1711.06104, 2017.

[20] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *ArXiv*, abs/1706.03825, 2017.

[21] K. Simonyan, A. Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2014.

[22] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 06–11 Aug 2017.

[23] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 2020.

[24] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[25] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Benjamin D. Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: an adaptive path method for removing noise. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5048–5056, 2021.

[26] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Vi'egas, and Michael Terry. Xrai: Better attributions through regions. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4947–4956, 2019.

[27] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025, 2022.

[28] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2203.12667*, 2022.

[29] Parsa Mahmoudieh, Deepak Pathak, and Trevor Darrell. Zero-shot reward specification via grounded natural language. In *ICLR 2022 Workshop on Generalizable Policy Learning in Physical World*, 2022.

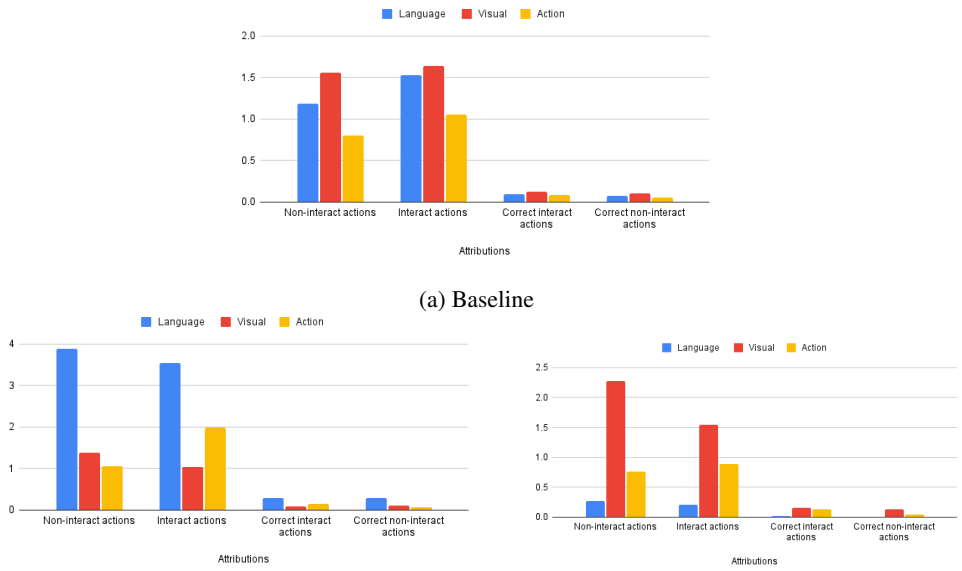
[30] Guorui Zhou and Guang Chen. Hierarchical latent semantic mapping for automated topic generation. *CoRR*, abs/1511.03546, 2015.

[31] Yichi Zhang and Joyce Chai. Hierarchical task learning from language instructions with unified transformers and self-monitoring. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Aug 2021.

A More Multimodal Attribution Analysis

We compare the multimodal attribution in terms of interact vs non-interact action 5, and how they may change with respect to % episode completion in Figures 6a, 6b and 6c. We also compare the biases in attributions to the modalities because of choices in model architecture and the way training is performed, i.e., dataset and learning techniques.

A.1 Interact vs non-interact action attributions



(a) Baseline

(b) MOCA

(c) E.T.

Figure 5: L2 norm of the attribution by grad $\odot$ input for interact and non-interact actions.

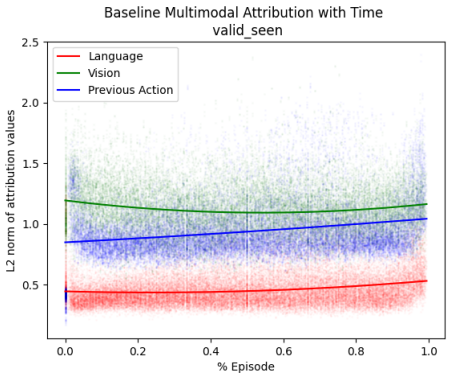
We analyze the L2 norm of the attribution over all the possible input-output pairs in valid seen and unseen trajectories and bin them in terms of interact (like pick, place, open, close, etc.) and non-interact (like move ahead, turn left, turn right, etc) actions in Fig. 5.

A.2 Attribution with respect to % episode completion

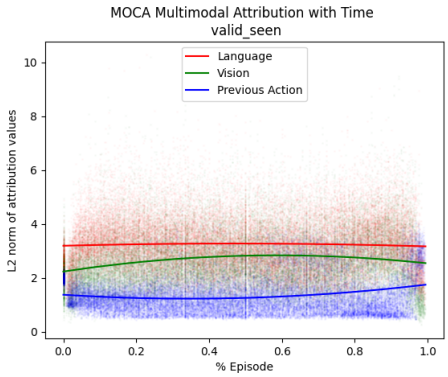
To compare the overall attribution over modality with respect to where the action is taken in terms of % episode length, we plot the attributions for the models trained on ALFRED task in Fig 6a, 6b, 6c and 6d to visualize how the attributions change during the episode for baseline, MOCA, ET and HiTUT respectively.

B What is the ideal target to estimate multimodal attribution?

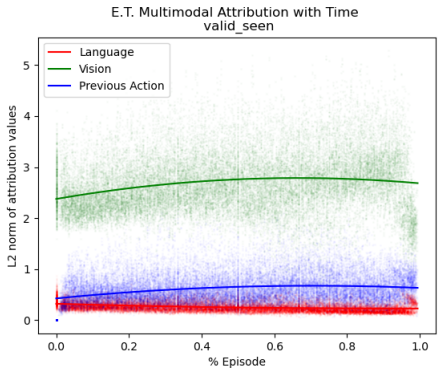
We compute the attribution of each modality/modality-specific feature with respect to the action predicted by the agent, that is f is a scalar value for the most likely action. Other values for function f can be other scalar values such as the loss used during training or the action taken by the expert at that particular timestep, or tensors, like the entire action space (since a policy would return logits for all actions in most cases) or the predicted interaction masks. Taking attributions with respect to the loss tells us what the agent should be looking at to better imitate the expert. But in terms of what features the agent is using to select an action gets conflated with the gradient of the loss with respect to that action. Taking attributions with respect to the expert action would result in an interpretation that would convey the attributions that would have led to the agent taking the expert action. Considering all the logits corresponding to the action choices makes sense for our purposes but there is a concern that not all model architectures allow us to backpropagate to inputs from all the possible discrete actions. Predicted interaction masks are a good indication of what the model is



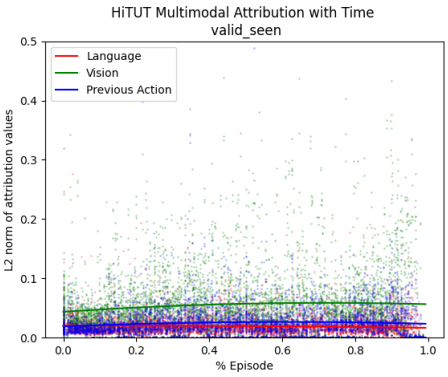
(a) Baseline [9]: Previous action attribution is higher than vision. Language is the lowest.



(b) MOCA [14]: Language attribution is the highest, especially towards the end of the trajectory.



(c) E.T. [15]: Visual attributions are the highest, followed by those for the previous action taken.



(d) HiTUT [31]: Overall attributions are low. Vision is higher, more towards the end.

Figure 6: Multimodal attributions in ALFRED policies over % episode completion for seen validation.

looking at while making an interact action. But this is a vision-biased attribution and also assumes that the model computes an interaction mask in the first place. Therefore, gradient with respect to the action taken is a better choice to interpret the policy decisions and it intuitively translates to the objective of filtering for inputs that gave rise to a certain decision.

C Limitations of the attribution methods

Attribution methods might be sensitive to the choice of pooling used. Here we defend our formulation, why L2 norm - because we want to compare policies that may use different dimension embeddings to represent an input. For example, policy 1 uses $d_1 = 128$ dim to represent word token embedding, policy 2 uses $d_2 = 768$ dim to represent the same. There are two factors here: first, in higher dimensions, the L1 norm gives a better distance estimate than the L2 norm. But, by the virtue of how the neural networks are initialized for stable training, the higher dimension will have a lower value per dimension. This means that the absolute attribution value in policy 1 should be scaled by its embedding size for a fair comparison to policy 2.