Attention to Action: Leveraging Attention for Object Navigation (Supplementary Materials)

Shi Chen
chen4595@umn.edu
Qi Zhao
qzhao@cs.umn.edu

Department of Computer Science and Engineering
University of Minnesota

The supplementary materials consist of additional experimental results and implementation details of comparative methods:

1. We investigate the usefulness of visual observations for object navigation with additional quantitative experiments (Section 1.1).
2. We study the effectiveness of joint optimization of perception and action by comparing the proposed method with its counterpart that uses independent optimization (Section 1.2).
3. We study the impacts of attention distribution on action prediction (Section 1.3)
4. We perform ablation analyses on the trainable balance factor to study the dynamics of attention-action relationship . (Section 1.4)
5. We perform ablation analyses to study the contributions of different components of our method. (Section 1.5)
6. We provide qualitative results of the proposed method (Section 1.6).
7. We present additional implementation details of different comparative methods discussed in the main paper, including the Baseline w/ coupling (Section 2.1) and the Dynamic Mapping method (Section 2.2).
8. We elaborate the details of our experiment settings, including training details of the agent (Section 2.3), target objects for training and evaluation (Section 2.4), action space (Section 2.5) and the classification of evaluation environments (Section 2.6).

1 Supplementary Results

1.1 Does Visual Observations Help Navigation?

One of the key challenges preventing the generalization of embodied agents is the visual variation of different environments. It leads to significant discrepancies between features extracted from visual observations of diverse environments, causing difficulties for agents to

	Known Semantics		Unknown Semantics	
	SR (%)	SPL (%)	SR (%)	SPL (%)
Baseline	51.31	18.84	34.80	7.80
Baseline w/o feat	46.18	13.34	36.86	7.52
ANA	**62.56**	**22.75**	**49.85**	**12.63**

Table 1: Comparison between the proposed ANA and different baselines. Following experimental settings in the main paper, experiments are conducted in unseen environments with known and unknown semantics. Best results are highlighted in bold.

determine their actions. To overcome the generalizability issue, in the main paper we propose to leverage a new attention mechanism that serves as a compact intermediate state bridging perception and action, distilling useful visual information while bypassing the direct use of visual features. In this subsection, we explore and report results with an extreme setting that completely removes the visual features extracted from the observations.

Specifically, we construct an additional model (Baseline w/o feat) that only utilizes the contextual information about object relationship for navigation. It is modified from the state-of-the-art Spatial Context [3] model (our Baseline) by removing the layers that take the visual features as inputs. We report comparative results in Table 1. Results show that (1) discarding the visual features leads to slight improvements over the Baseline on the Successful Rate (SR) in environments with unknown semantics but significantly decreases the performance on the other evaluation metrics. It shows that while having a heavy reliance on the visual features from the visual observations limits generalization, completely removing them does not help much. (2) Instead, by distilling useful information from visual observations into attention distribution, the proposed ANA shows a significantly improved performance across various types of unseen environments.

1.2 Does Joint Optimization of Attention and Action Help Navigation?

A key component of the proposed method is the joint optimization of attention and action through a consistent action space. In Section 4.2 of the main paper, we show that the joint optimization is more advantageous than implicitly optimization without coupling attention and action (Implicit Optimization, equivalent to Baseline w/ proposed attention). To gain more insights of the paradigm, we further compare it with a self-supervised learning method that also explicitly optimizes attention and action but with separate training objectives (Independent Optimization).

Specifically, this compared method derives attention ground truth based on the predicted action, and encourages the model to look at regions related to the action. It adaptively constructs the attention ground truth α_t^{GT} by multiplying the predicted action likelihood Act_t with spatial masks $M \in R^{k\times 7\times 7}$ that indicate the important regions for different candidate actions:

$$\alpha_t^{GT} = \frac{1}{Z}\sum_k Act_t \cdot M \tag{1}$$

where k represents the number of actions, Z is the sum of attention values for normalizing the ground truth. The masks are initialized based on the same heuristic used in our method, and also trained end-to-end for data-driven refinement and improved flexibility.

Following a similar intuition as our method ANA, we encourage the agent to focus on

	Known Sem.		Unknown Sem.	
	SR (%)	SPL (%)	SR (%)	SPL (%)
Implicit Optimization	54.02	17.83	42.58	11.12
Independent Optimization	54.58	17.91	40.67	7.41
ANA	**62.56**	**22.75**	**49.85**	**12.63**

Table 2: Comparison between methods with different optimization strategies. Best scores are highlighted in bold.

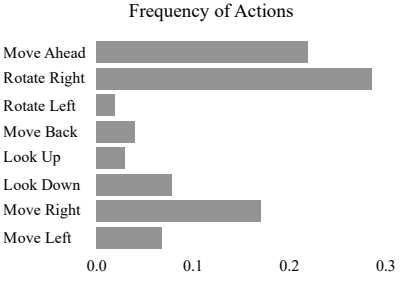

Figure 1: Frequency of actions performed by the agent.

the important regions associated with the predicted action:

$$L_{att} = -\log(\sum_{w,h} \alpha_t^{GT} \cdot \alpha_t) \tag{2}$$

The attention loss L_{att} is linearly combined with the navigation objective L_{nav}:

$$L = L_{nav} + \gamma L_{att} \tag{3}$$

where γ is the balance factor (we empirically define $\gamma = 0.2$ to maintain a consistent scale between the two objectives).

As shown in Table 2, while Independent Optimization slightly outperforms Implicit Optimization in environments with known semantics, it leads to a considerable drop of performance under the setting with unknown semantics, where the integration of perception and action plays a more significant role. Without explicitly coupling attention and action for a joint optimization, the agent does not learn well the relationship between attention and action, which is made worse by the unbalanced frequency of candidate actions, *i.e.*, the agent tends to optimize dominant actions (see Figure 1). Differently, by explicitly coupling attention with action, our method is less prone to the unbalanced frequency and able to generalize across various types of unseen environments.

1.3 Does Where to Look Affect Where to Move?

While many efforts have been placed on improving the overall model performance with attention, less reported the contribution of attention on each prediction. Moreover, several recent studies [5, 9] in the Natural Language Processing (NLP) community point out the lack of explainability of attention, and show that altering the attention of language models does not have significant impacts on their predictions. In this section, we focus on the influence of attention on predicting actions, and investigate if the same issue exists in object navigation.

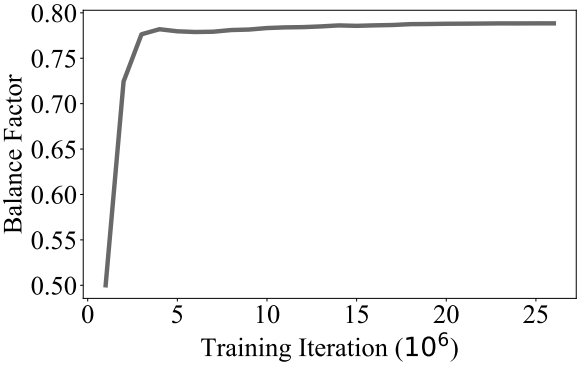

Figure 2: Adaptive values of the balance factor β for different training iterations.

Following [5], we randomly permute the attention computed by the fully trained models, and record the corresponding changes in action. We perform the experiment on both the proposed ANA with the new attention, and a baseline model with conventional attention (*i.e.*, Baseline w/ soft attention). Specifically, the Baseline w/ soft attention model follows the typical designs of attentive models [1, 2, 4, 6, 8], and incorporates conventional soft attention with the Spatial Context [3] baseline. By averaging results across different episodes, we find that in 51.66% of the times Baseline w/ soft attention model changes its action. It suggests that the influence of attention on the final predictions is task-dependent. Compared to NLP tasks that require an understanding of a long sequence of words, object navigation with a few regions of interest tends to have a stronger reliance on attention. Moreover, with the proposed attention that explicitly integrates perception and action, ANA shows an increased probability of 70.26%. It indicates that by coupling attention with action, our attention plays an essential role in guiding the action of agent and boosting the navigating performance.

1.4 Ablation Study on the Balance Factor

Our method takes advantage of a trainable balance factor β (see Equation 8) to determine the contribution of information from different aspects, *i.e.,* the alignment between attention and spatial masks for action Act_t^{α} and the cooperation of attention and contextual information Act_t^{c}. In Figure 2, we visualize values of the balance factor for different training iterations. The results show that β increases drastically at the beginning, during which the agent quickly develops its initial navigation policy. After that, it maintains a slowly increasing trend throughout the rest of the learning process, as the agent iteratively refines its policy. The overall trend of the balance factor suggests that our method integrates perception and action in a progressive manner. As the agent develops better understanding of the task, coupling attention with action plays a more and more important role and leads to more accurate action planning.

1.5 Ablation Study on Different Components

The proposed method jointly considers contextual information and attention distribution. By mapping attention to the action space (*i.e., Act_t^{α}* in Equation 3 of the main paper) and learning discriminative features from both attention and contextual information (*i.e., Act_t^{c}*

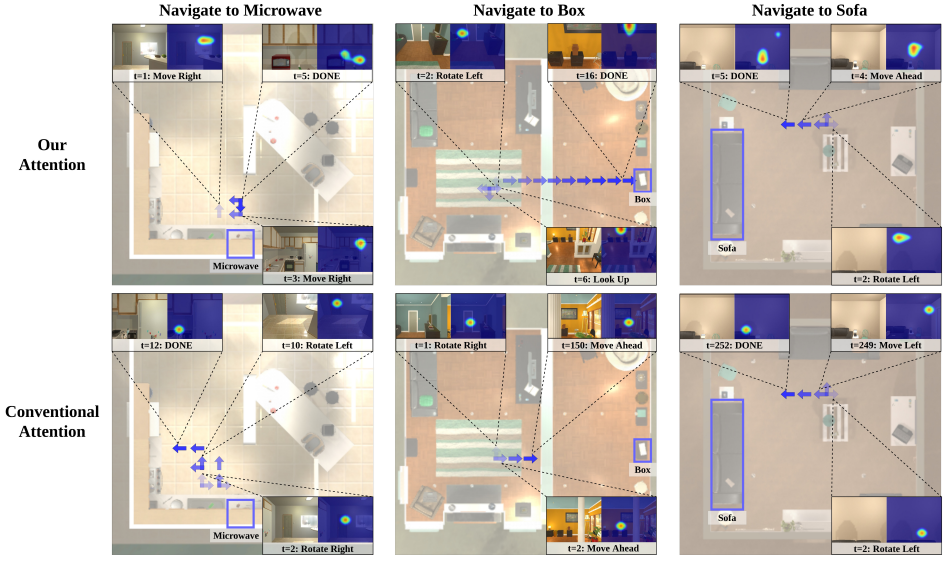

Figure 3: Qualitative comparison between agents with the conventional attention (Baseline w/ soft attention) and the proposed attention (ANA). The trajectories of agents are are denoted with blue arrow, whose opacity represents the order of actions performed by agents (the higher the opacity, the later the action takes place). The final targets are highlighted with blue bounding boxes.

	Known Sem.		Unknown Sem.	
	SR (%)	SPL (%)	SR (%)	SPL (%)
ANA	**62.56**	**22.75**	**49.85**	**12.63**
ANA w/o Act_t^c	55.70	20.59	34.96	7.92
ANA w/o Act_t^α	54.02	17.83	42.58	11.12

Table 3: Ablation results for different components. Best scores are highlighted in bold.

in Equation 2 of the main paper), it is able to effectively navigate across various visual environments. In this subsection, we perform an ablation study to study the contribution of the two components. As shown in Table 3, dropping either component leads to a considerable loss of performance, which highlights the complementary role of the components and the integral design of our method.

1.6 Additional Qualitative Results

This subsection provides qualitative analyses between conventional attention and the proposed attention. In Figure 3, we visualize the navigation episodes of agents equipped with the two different attention. Results show that conventional feature aggregation based attention (second row of Figure 3) commonly fails to capture the regions of interest, and focuses on the background instead (*e.g.,* attention for $t = 2, 10$ in the first episode and all three steps in the third episodes). As a result, the agent has difficulty navigating to the targets, and tends to collide with different obstacles (*e.g.,* the second and third episodes). On the contrary, in

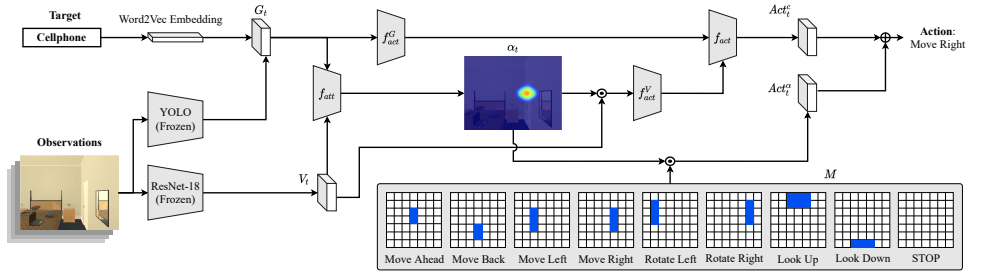

Figure 4: Architecture of the Baseline w/ coupling method, a comparative method we use in the main paper.

the proposed method (ANA, first row of Figure 3), our new attention highlights the potential directions of the final targets, and is closely correlated with the action. With the guidance of the proposed attention, our method is able to efficiently locate the targets without running into collision.

2 Supplementary Method

2.1 Coupling Conventional Soft Attention with Action

This subsection elaborates the implementation details of the Baseline w/ coupling model used as a comparative method in the main paper. The model (shown in Figure 4) shares a similar architectural design with the proposed method, with the key difference being the use of attention, *i.e.*, it uses conventional soft attention for aggregating visual features instead of our proposed attention as the intermediate state.

Specifically, it first encodes the visual features V_t and context grid G_t independently, and then concatenates the encoded features $\hat{V}_t$ and $\hat{G}_t$ to compute the visual attention α_t:

$$\alpha_t = \sigma(f_{att}([\hat{V}_t; \hat{G}_t])) \tag{4}$$

where f_v and f_g denote the layers for processing visual features and context grid, f_{att} corresponds to layers for computing the attention. $[\cdot]$ represents concatenation of features, and σ is the softmax activation function.

After obtaining the attention, we follow the conventional feature aggregation scheme [1, 9, 10], and leverage it to adaptively determine the contribution of visual features:

$$V_t^{\alpha} = \sum_{w,h} \alpha_t \cdot V_t \tag{5}$$

where w and h represent the spatial dimension, $\cdot$ denotes the Hadamard product. The attended visual features V_t^{α} are then concatenated with the features derived from the context grid for predicting the unnormalized action likelihood Act_t^c:

$$Act_t^c = f_{act}([f^V_{act}(V_t^{\alpha}); f^G_{act}(G_t)]) \tag{6}$$

where f^V_{act} and f^G_{act} are fully-connected layers for encoding the attended visual features and context grid, and f_{act} denotes fully-connected layers for deriving the action.

	Known Targets	Unknown Targets
Living Room	Pillow, Laptop, Television, Garbage Can, Bowl	Sofa, Box, Table Top
Bathroom	Sink, Toilet Paper, Soap Bottle, Light Switch	Toilet, Towel
Kitchen	Toaster, Microwave, Fridge, Coffee Machine, Garbage Can, Bowl	Mug, Pot, Cup
Bedroom	House Plant, Lamp, Book, Alarm Clock	Mirror, CD, Cellphone

Table 4: Target object categories for different room types.

Similar to the proposed method, Baseline w/ coupling also couples attention with action, and determines the final action with the consideration of the alignment between attention and action templates Act_t^{α}:

$$Act_t^{\alpha} = \sum_{w,h} M \cdot \alpha_t \tag{7}$$

$$Act_t = \sigma(Act_t^c + \beta \cdot Act_t^{\alpha}) \tag{8}$$

where M is the spatial masks discussed in the main paper, β is the trainable balance factor, and σ is the softmax activation function.

Despite integrating perception and action by mapping attention to the action space, as the conventional feature-aggregation based attention is not directly correlated with action, Baseline w/ coupling does not bring reasonable improvements and is significantly outperformed by our method with the newly proposed attention.

2.2 Dynamic Mapping for Integrating Perception and Action

In the main paper, we analyze the effectiveness of different spatial masks through comparative experiments. In this subsection, we elaborate the detailed design of the Dynamic Mapping method used in the analyses. Different from our method that applies the same set of masks across different environments, it dynamically determines the masks for each time step based on visual observations and contextual information between objects.

The model shares a similar design as the proposed ANA, except having an additional dynamic mapping module for computing the spatial masks. Given the visual features V_t extracted from the four most recent observations and the context grid G_t encoding object relationship, the module first independently encodes the two types of features with a sequence of convolutional layers and pooling layers. After obtaining the encoded features V_t^M and G_t^M of a consistent spatial dimension, the spatial masks M_t for the current time step t are computed as:

$$M_t = \sigma(f_M([V_t^M; G_t^M])) \tag{9}$$

where f_M is a convolutional layer for computing the masks, $[\cdot]$ denotes the concatenation operation, and σ is the Sigmoid activation function.

While taking into account the dynamics of environments, due to the absence of prior knowledge about the relationship between attention and candidate actions, the Dynamic Mapping method is not as effective as the proposed ANA that incorporates both prior knowledge and environment dynamics.

2.3 Optimization

Our agent is trained under the Asynchronous Advantage Actor-Critic (A3C) [7] algorithm using 6 threads, with a mixture of rewards defined in [3]. We train the agent with 25 millions

Figure 5: Examples of environments in w/ obstacles and Obstacle-free groups.

	Environment ID
w/ obstacles	22, 30, 224, 226, 228, 230, 322, 326, 330, 426
Obstacle-free	24, 26, 28, 222, 324, 328, 422, 424, 428, 430

Table 5: Environment IDs for environments in different groups.

iterations using the RMSProp optimizer. The learning rate is initialized as 7×10^{-4}, and decreases linearly till 0 at the last iteration.

2.4 Target Objects

The target objects in our experiments are determined based on previous state-of-the-art [3, 11]. For the setting with **Unseen environments with known semantics**, the target objects for training and evaluation are consistent, *i.e.,* Known Targets in Table 4. In terms of setting with **Unseen environments with unknown semantics**, the target objects for evaluation (Unknown Targets in Table 4) are determined by selecting objects closest to training targets (Known Targets) on the word embedding space. This allows us to evaluate the effectiveness of agents on navigating to semantically relevant objects.

2.5 Action Space

Following [3], the action space in our experiments consists of 9 unique actions, which are defined as follows:

- **Move Ahead, Move Back, Move Left, Move Right**: these actions will move the agent to the corresponding direction for a single step (*i.e.,* 0.5 meter).
- **Rotate Left, Rotate Right**: the two actions that rotate the agent to the corresponding directions for 90 degrees.
- **Look Up, Look Down**: these two actions will not move the agent, but instead will tilt its camera up or down for 30 degrees.
- **STOP**: a special action for terminating an episode.

2.6 Classification of Environments

To study the influence of environments on the relationship between attention and action, in the main paper we performed an analysis with two groups of environments (*i.e.,* w/ obstacles and Obstacle-free). This subsection visualizes exemplar environments for both groups (Figure 5), and lists the complete assignment for all evaluation environments (Table 5). The classification was determined based on examining the amount of free-space in an environment from a top-down view.

Acknowledgements

This work is supported by NSF Grants 1908711 and 1849107.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.

[2] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2018.

[3] Raphael Druon, Yusuke Yoshiyasu, Asako Kanezaki, and Alassane Watt. Visual object search by learning spatial context. *IEEE Robotics and Automation Letters*, 5(2):1279–1286, 2020.

[4] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *Conference on Neural Information Processing Systems*, page 3318–3329, 2018.

[5] Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3543–3556, 2019.

[6] Anthony Manchin, Ehsan Abbasnejad, and Anton van den Hengel. Reinforcement learning with attention that works: A self-supervised approach. In Tom Gedeon, Kok Wai Wong, and Minho Lee, editors, *International Conference on Neural Information Processing*, pages 223–230, 2019.

[7] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Tim Harley, Timothy P. Lillicrap, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, page 1928–1937, 2016.

[8] Alexander Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, and Danilo Jimenez Rezende. Towards interpretable reinforcement learning using attention augmented

agents. In *Conference on Neural Information Processing Systems*, pages 12329–12338, 2019.

[9] Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, 2019.

[10] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6622–6631, 2019.

[11] Mitchell Wortsman, Kiana Ehsani, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Learning to learn how to learn: Self-adaptive visual navigation using meta-learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6743–6752, 2019.