Supplementary Material: Spatiotemporal Attacks for Embodied Agents

Aishan Liu[1], Tairan Huang[1], Xianglong Liu[1,2]*, Yitao Xu[1], Yuqing Ma[1],
Xinyun Chen[3], Stephen J. Maybank[4], and Dacheng Tao[5]

[1] State Key Laboratory of Software Development Environment,
Beihang University, China
[2] Beijing Advanced Innovation Center for Big Data-Based Precision Medicine,
Beihang University, China
[3] UC Berkeley, USA
[4] Birkbeck, University of London, UK
[5] UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering,
The University of Sydney, Australia

1 More Details of Experimental Settings

1.1 Generalization Ability of the Attack

We provide the details of our experimental settings in Section 5.6. We use 99 questions of 5 houses for both of the following 2 settings.

Different questions. For each question, we generate spatiotemporal perturbations based on the current question, and then evaluate another question given the scenes with the same perturbed object.

Different starting points. For each question, we randomly sample another question, and then use its starting point as the initilization point for the agent to answer the current question. The average distance change for the starting point is 2.47 (maximum is 9.52, minimum is 0.83). Among the 99 new starting points, 45.45% of them are in the different rooms.

1.2 Improving Agent Robustness with Adversarial Training

We provide the details of our experimental settings in Section 5.7, where we evaluate the effectiveness of adversarial training for T_{-10} setting.

Training. We use the SGD optimizer for adversarial training, with a learning rate of 0.001. Following [1], both the QA and NAV modules are trained for 300 epochs. In each training batch, we generate one perturbed scene (either adding the adversarial perturbation or the Gaussian noise) for each clean scene, so that the numbers of clean scenes and perturbed scenes are the same per batch, *i.e.*, 4 clean scenes and 4 perturbed scenes per batch in our experiments. The magnitude of adversarial perturbations is 32/255. For Gaussian noises, we choose the maximum noise severity level following [3], and set the mean to be 0, the standard deviation to be 0.38. The other settings are the same as that in Section 5.3.

Testing. For evaluation, we use the same approaches as for training to add either adversarial perturbations or Gaussian noises to the chosen 3D objects.

1.3 Perceptual Studies via Amazon Mechanical Turk (AMT)

We design a perceptual study on AMT to figure out which features are more sensitive and attractive for human predictions, *i.e.*, shape or texture. For each question, the participants need to select the correct category of the object in the picture. We do not set any time limit for the responses.

In total, we collect 30 objects in different scenes, each of which is perturbed on shape and texture, respectively. Thus, we have a total of 60 questions, namely 60 Human Intelligence Tasks (HITs). For each HIT, we make 10 assignments, *i.e.*, each HIT will be answered by 10 different human workers. As a result, we finally collect 600 responses for our perceptual study.

For fair comparisons, we use the same hyper-parameters for shape and texture attacks. We limit the overall perturbation magnitude to 32/255, as in other settings.

2 Additional Experimental Results

In this section, we provide more experimental results.

2.1 Texture v.s. Shape

In this section, we study the importance of texture and shape for model predictions. For a fair comparison, we set the same constraint of perturbation magnitude for both texture and shape attacks. According to the accuracy of the texture attack (4.26%) and shape attack (27.14%) in the T_{-30} task, perturbing textures is far more effective than perturbing shapes. A question emerges: *Which is more important for model prediction, texture or shape?*

A recent study [2] demonstrated that CNNs are strongly biased towards recognizing textures. Compared to long-range dependencies encoded in the shapes of objects, standard CNNs prefer local textures [4]. Thus, it is not uncommon to see that the agent is more likely to make errors when 3D object textures are adversarially perturbed.

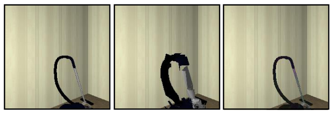


Fig. 1. Visualization of scene perturbed on different physical parameters. From left to right: clean, shape attacks, and texture attacks.

Since deep learning prefers textural information when making decisions, it is worth studying which features humans find more beneficial. As a preliminary step, we examined which features are more sensitive for human predictions with a user study conducted on the Amazon Mechanical Turk (AMT). With each object adversarially perturbed in texture and shape (c.f. Figure 1), participants were asked to assign those adversarial objects to one of five classes (the ground-truth class, the top-3 adversarial target classes, and "none of the above"). Our results showed that the classification accuracy for adversarial texture manipulation (83.3%) was higher than that for shape (32.7%). It indicates that shape is a more sensitive parameter for human predictions compared to texture. This is obvious since people are more likely to recognize a table with different textures rather than a table made out of wood but showing a strange shape.

In conclusion, embodied agents trained upon most current strategies are more sensitive to texture rather than shape. It is in stark contrast to humans and reveals fundamental differences in classification strategies between humans and machines. Therefore, to bridge the gap between human perception and embodied perception, it is important to train agents that can better capture shape-based features. Could we obtain stronger policies for agents if we train them with shape-based adversarial perturbations? We put it as future work.

2.2 Attention similarity

To understand the transferability of attacks between different models, we investigate their attention correlation. We first visualize the attention map of the last 5 views using PACMAN and VIS-VGG in Figure 2, and we observe that the attention zones highlight similar context of the scenes for prediction. Moreover, we compare the top-3 important views between PACMAN and VIS-VGG on 32 questions, and we find that 83.33% of the included views are the same for both models. Such attention similarities between different models could facilitate the transferability of black-box attacks.

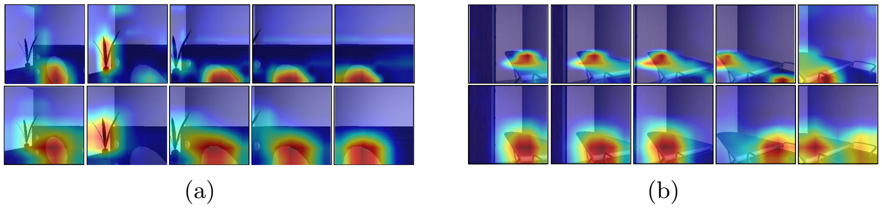
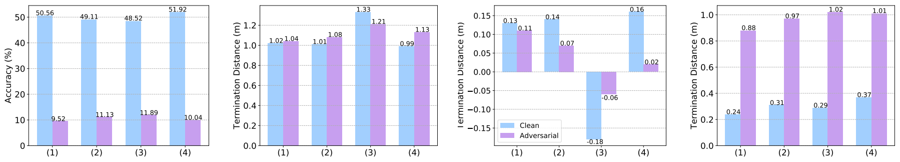

(a) (b)

Fig. 2. The attention maps of different models. In both scenes (a) and (b), the first line presents the attention maps of PACMAN, and the second line presents those of VIS-VGG. The attention zones highlight similar context of the scenes for prediction.

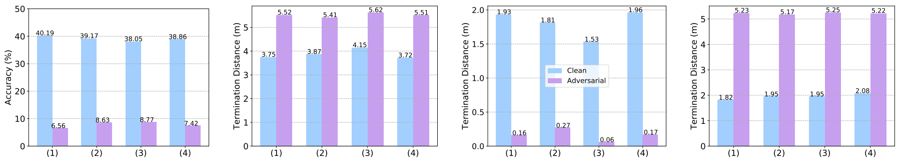
2.3 Transfer Attack onto a non-differentiable Renderer

Here, we present more experimental results of transferring our generated spatiotemporal perturbations to attack a non-differentiable renderer for EQA tasks. In addition to the results of T_{-30} task discussed in Section 5.5, we further show the results of T_{-10} and T_{-50} tasks in Figure 3 and 4, respectively. Again, we observe that our attacks transfer to the non-differentiable renderer.



(a) Accuracy (b) d_T (c) d_Δ (d) d_{min}

Fig. 3. Transfer attack on a non-differentiable renderer for task T_{-10}. Methods (1) to (4) represent PACMAN, NAV-GRU, NAV-React, and VIS-VGG, respectively.

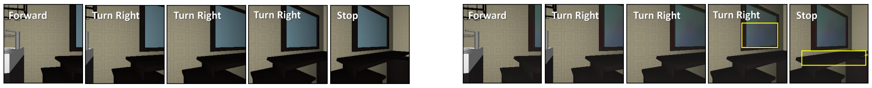


(a) Accuracy (b) d_T (c) d_Δ (d) d_{min}

Fig. 4. Transfer attack on a non-differentiable renderer for task T_{-50}. Methods (1) to (4) represent PACMAN, NAV-GRU, NAV-React, and VIS-VGG, respectively.

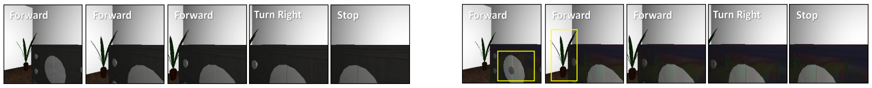
2.4 Sample Adversarial Attacks for Question Answering and Visual Recognition

In this section, we show more examples of adversarial scenes generated using our attack framework. Figures 5, 6, 7, and 8 illustrate some examples of our adversarial attacks for question answering. All of these questions are answered correctly by agents in clean scenes, but wrongly in corresponding adversarial scenes. Examples for visual recognition are shown in Figures 9, 10 and 11. All of these objects are classified correctly by agents in clean scenes, but wrongly in corresponding adversarial scenes.



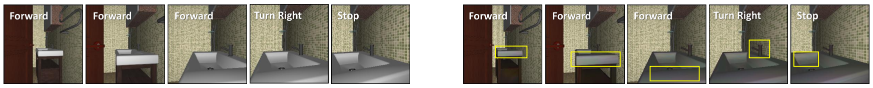
(a) Clean scene (b) Adversarial scene

Fig. 5. Given the question "*What color is the mirror?*", we show the last 5 views of the agent for EQA in the same scene with and without adversarial perturbations. The contextual objects perturbed including table and mirror. The ground truth prediction is "white". The agent gives the wrong answer "black" to the question.



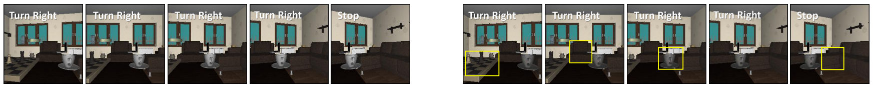
(a) Clean scene (b) Adversarial scene

Fig. 6. Given the question "*What color is the shoes cabinet?*", we show the last 5 views of the agent for EQA in the same scene with and without adversarial perturbations. The contextual objects perturbed including cabinet and plant. The ground truth prediction is "brown". The agent gives the wrong answer "yellow" to the question.



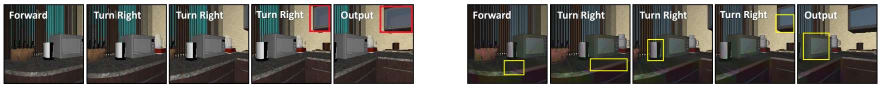
(a) Clean scene (b) Adversarial scene

Fig. 7. Given the question "*What room is the sink located in?*", we show the last 5 views of the agent for EQA in the same scene with and without adversarial perturbations. The contextual objects perturbed including sink and water tap. The ground truth prediction is "bathroom". The agent gives the wrong answer "kitchen" to the question.



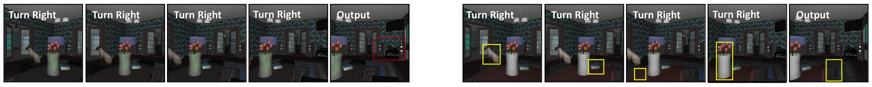
(a) Clean scene (b) Adversarial scene

Fig. 8. Given the question "*What room is the cup located in?*", we show the last 5 views of the agent for EQA in the same scene with and without adversarial perturbations. The contextual objects perturbed including chessboard, cup, and sofa. The ground truth prediction is "living room". The agent gives the wrong answer "bedroom" to the question.



(a) Clean Scene (b) Adversarial Scene

Fig. 9. The last 5 views of the agent for EVR in the same scene with and without adversarial perturbations. The contextual objects perturbed are table, kettle, microwave, and cabinet. After the adversarial attack, the agent fails to recognize the cabinet in subfigure (b). Red boxes indicate the bounding box for object detection and yellow boxes show the adversarially perturbed texture regions.



(a) Clean scene (b) Adversarial scene

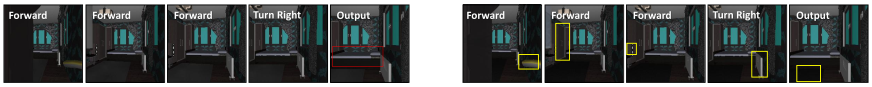
Fig. 10. The last 5 views of the agent for EVR in the same scene with and without adversarial perturbations. The contextual objects perturbed are dog, book, desk, vase, and carpet. After the adversarial attack, the agent fails to recognize the stereoset in subfigure (b). Red boxes indicate the bounding box for object detection and yellow boxes show the adversarially perturbed texture regions.

2.5 Sample Attention Maps of the Agent for Question Answering

As shown in Figures 12, 13, 14, and 15, we provide more visualization of the egocentric views and corresponding attention maps when agents answer questions. We can observe that the agents use clues from contextual objects to answer locational and compositional questions while mainly focus on target objects when predicting their colors.

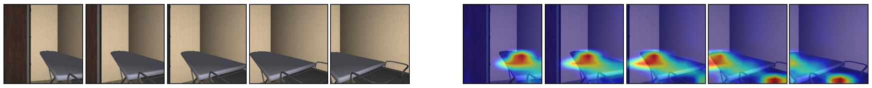
2.6 Sample Adversarial Attacks for Navigation

In this section, we show more examples of adversarial scenes for navigation in Figures 16, 17, and 18. Agents navigate correctly to the end in all of the clean scenes, but stop ahead of time in corresponding adversarial scenes.

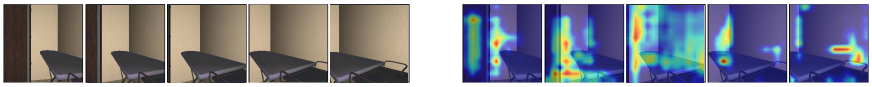


(a) Clean scene (b) Adversarial scene

Fig. 11. The last 5 views of the agent for EVR in the same scene with and without adversarial perturbations. The contextual objects perturbed are sofa, cabinet, stereo set, heating, and carpet. After the adversarial attack, the agent fails to recognize the bed in subfigure (b). Red boxes indicate the bounding box for object detection and yellow boxes show the adversarially perturbed texture regions.

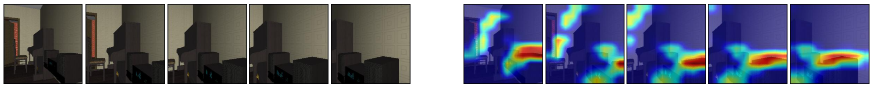


(a) Clean scene (b) Attention in clean scene

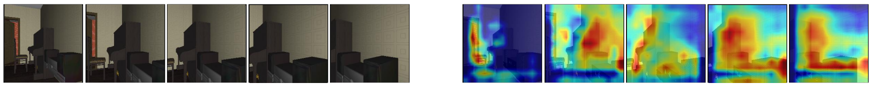


(c) Adversarial scene (d) Attention in adversarial scene

Fig. 12. Egocentric views and corresponding attention maps when the agent answers the question, "*What color is the ironing board?*". The agent mainly focuses on the target object when predicting its color in the clean scene (subfigure (a) and (b)). The adversarial scene and corresponding attention maps are shown in subfigure (c) and (d). The ground truth prediction is "white". The agent gives the wrong answer "brown" to the question.

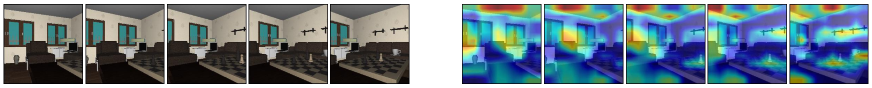


(a) Clean scene (b) Attention in clean scene

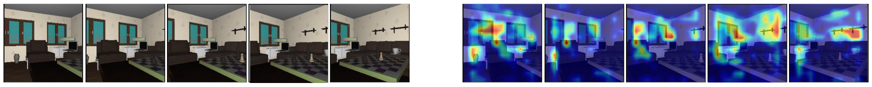


(c) Adversarial scene (d) Attention in adversarial scene

Fig. 13. Egocentric views and corresponding attention maps when the agent answers the question, "*What color is the stereo set?*". The agent mainly focuses on the target object when predicting its color in the clean scene (subfigure (a) and (b)). The adversarial scene and corresponding attention maps are shown in subfigure (c) and (d). The ground truth prediction is "black". The agent gives the wrong answer "white" to the question.

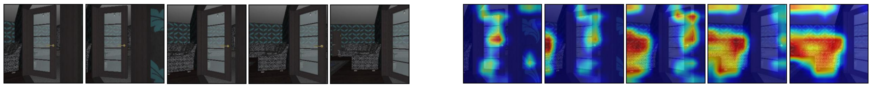


(a) Clean scene (b) Attention in clean scene

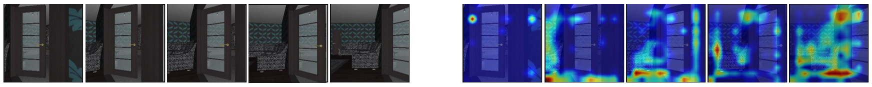


(c) Adversarial scene (d) Attention in adversarial scene

Fig. 14. Egocentric views and corresponding attention maps when the agent answers the question, "*What room is the chessboard located in?*". The agent uses clues from contextual objects to answer locational and compositional questions in the clean scene (subfigure (a) and (b)). The adversarial scene and corresponding attention maps are shown in subfigure (c) and (d). The ground truth prediction is "living room". The agent gives the wrong answer "bathroom" to the question.

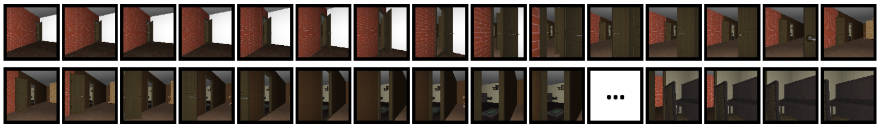


(a) Clean scene (b) Attention in clean scene

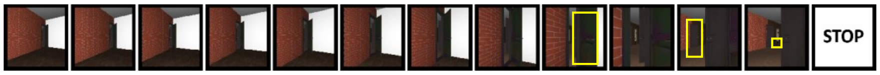


(c) Adversarial scene (d) Attention in adversarial scene

Fig. 15. Egocentric views and corresponding attention maps when the agent answers the question, "*What room is the toy located in?*". The agent uses clues from contextual objects to answer locational and compositional questions in the clean scene (subfigure (a) and (b)). The adversarial scene and corresponding attention maps are shown in subfigure (c) and (d). The ground truth prediction is "living room". The agent gives the wrong answer "kitchen" to the question.

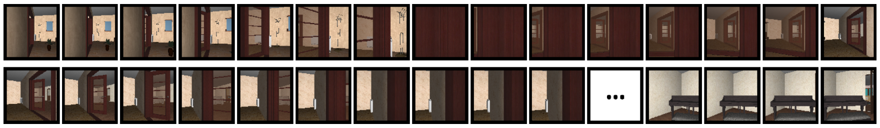


(a) Clean scene

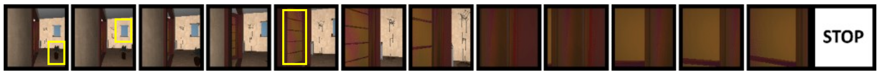


(b) Adversarial scene

Fig. 16. Egocentric views of the agent in the same scene with and without adversarial perturbations. As shown in subfigure (b), we perturb the textures of two doors and the rug. The agent stops at the 5-*th planner* step.

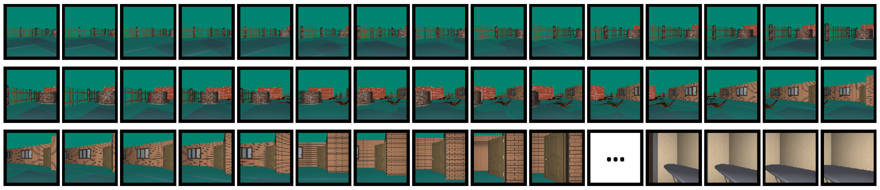


(a) Clean scene

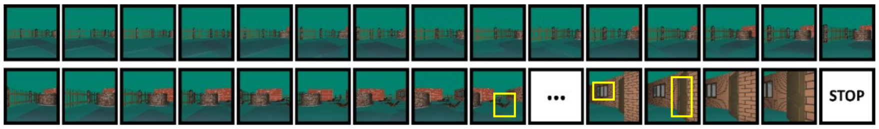


(b) Adversarial scene

Fig. 17. Egocentric views of the agent in the same scene with and without adversarial perturbations. As shown in subfigure (b), we perturb the textures of the plant, the door, and the window. The agent stops at the 4-*th planner* step.



(a) Clean scene



(b) Adversarial scene

Fig. 18. Egocentric views of the agent in the same scene with and without adversarial perturbations. As shown in subfigure (b), we perturb the textures of the bench, the door, and the window. The agent stops at the 12-*th planner* step.

References

1. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., Batra, D.: Embodied question answering. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
2. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
3. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2019)
4. Zhang, T., Zhu, Z.: Interpreting adversarially trained convolutional neural networks. arXiv preprint arXiv:1905.09797 (2019)