DialMAT: Dialogue-Enabled Transformer with Moment-Based Adversarial Training

DialFRED Challenge Winner

Kanta Kaneda*, Ryosuke Korekata*, Yuiga Wada*, Shunya Nagashima*, Motonari Kambara,
Yui Iioka, Haruka Matsuo, Yuto Imai, Takayuki Nishimura, and Komei Sugiura (Keio University) *Equal contribution

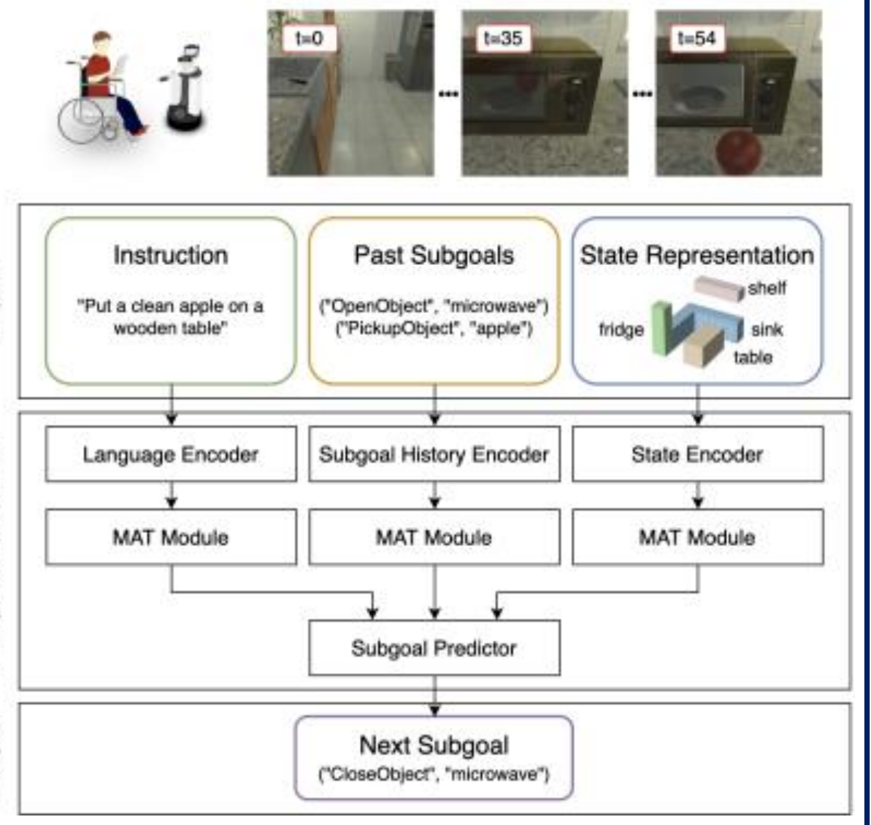
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

Major challenges of existing benchmarks

- Resolving ambiguities in open-vocabulary instructions
- Recovering from failed actions

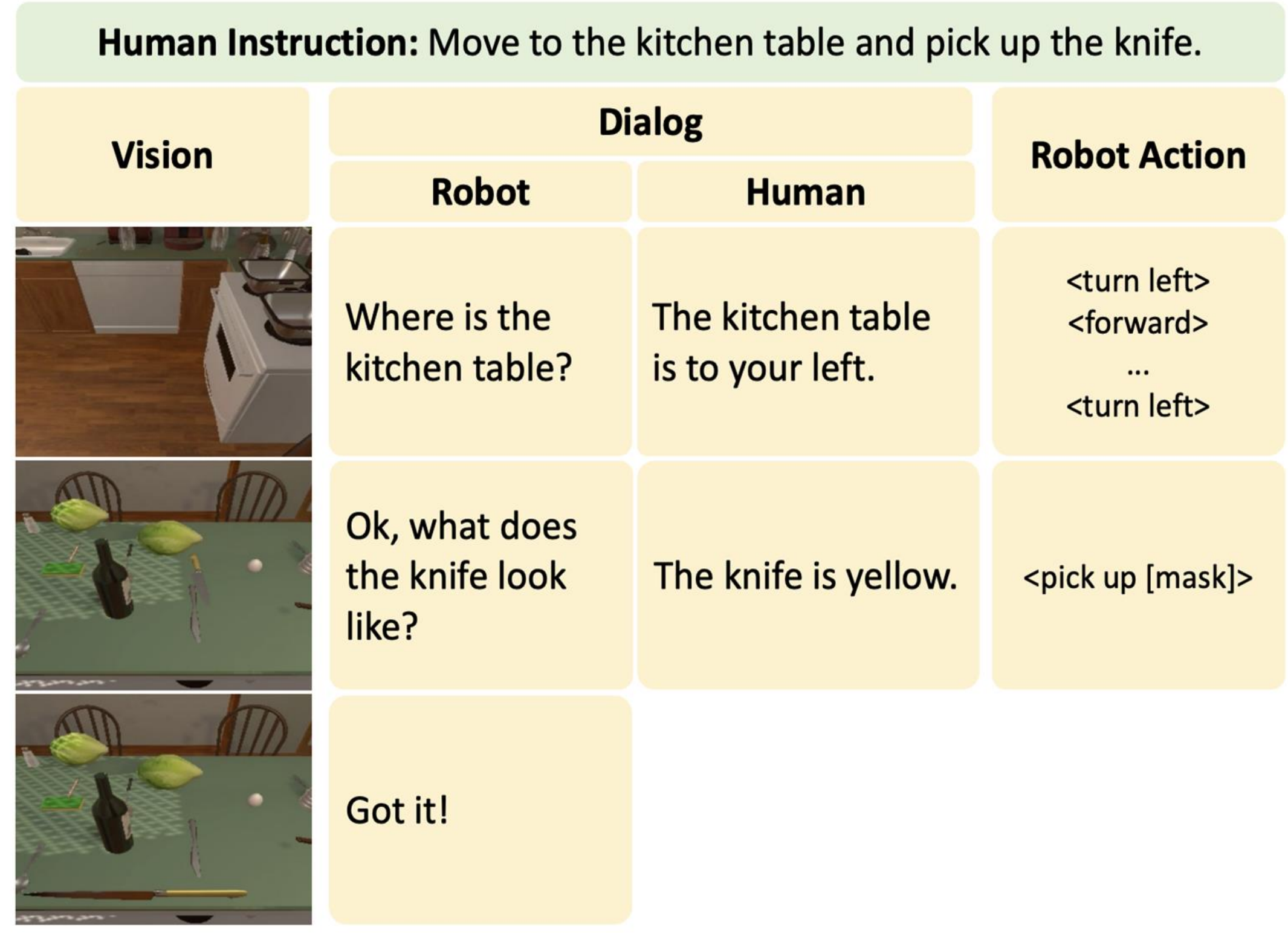
REVERIE Challenge 2022



[Ishikawa+, ICPR22]

DialFRED Task

- The task of embodied instruction
- Setting: an agent can actively ask questions to the human user
- e.g.,) Where is the knife?



Human Instruction: Move to the kitchen table and pick up the knife.

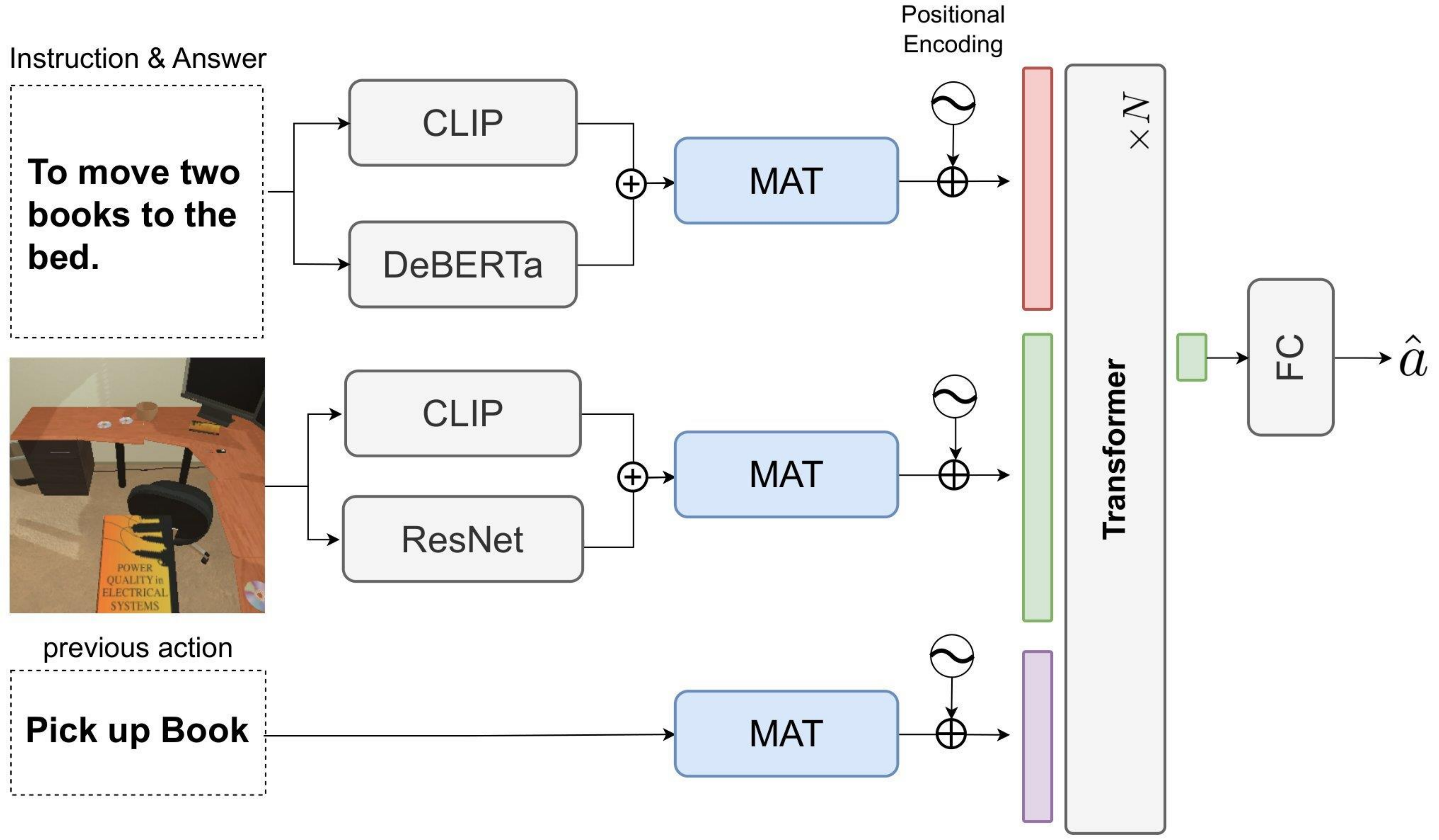
Vision	Dialog		Robot Action
	Robot	Human	
	Where is the kitchen table?	The kitchen table is to your left.	<turn left> <forward> ... <turn left>
	Ok, what does the knife look like?	The knife is yellow.	<pick up [mask]>
	Got it!		

Related Work

Task	Method / Benchmark
ALFRED [Shridhar+, CVPR20]	Prompter [Inoue+, 22], FILM [Min+, ICLR22], HLSM-MAT [Ishikawa+, ICPR22], E.T. [Pashevich+, ICCV21]
Object Navigation with dialogue	DialFRED [Gao+, RA-L22], TEACh [Padmakumar+, AAAI22], Vision-and-Dialog Navigation [Thomason+, CoRL19]

Methods

- **Moment-based Adversarial Training (MAT) [Ishikawa+, ICPR22]**
  - Add adversarial perturbation to the embedding spaces of language, image and action
- **A crossmodal parallel feature extraction mechanism using foundation models**



Step 1: Add adversarial perturbation to the embedding spaces

$$\nabla_{\boldsymbol{\delta}} E(\boldsymbol{\delta}) = \frac{\partial E}{\partial \boldsymbol{\delta}}$$

$$\boldsymbol{m}_t = \rho_1 \boldsymbol{m}_{t-1} + (1-\rho_1)\nabla_{\boldsymbol{\delta}} E(\boldsymbol{\delta}_t),$$

$$\boldsymbol{v}_t = \rho_2 \boldsymbol{v}_{t-1} + (1-\rho_2)(\nabla_{\boldsymbol{\delta}} E(\boldsymbol{\delta}_t))^2$$

Step 2: Update the perturbation

$$\hat{\boldsymbol{m}}_t = \frac{\boldsymbol{m}_t}{1-(\rho_1)^t}, \ \hat{\boldsymbol{v}}_t = \frac{\boldsymbol{v}_t}{1-(\rho_2)^t},$$

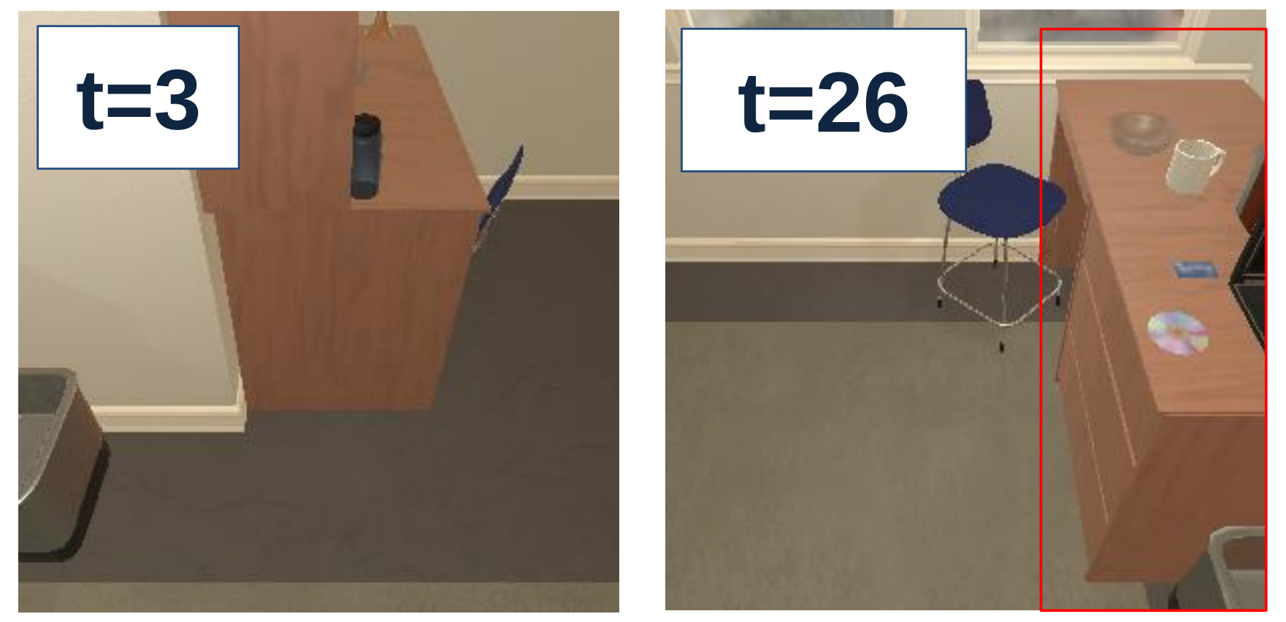
$$\Delta \boldsymbol{\delta}_t = \eta \frac{\hat{\boldsymbol{m}}_t}{\sqrt{\hat{\boldsymbol{v}}_t + \varepsilon}}$$

Results

- Divide the valid_unseen set
  - (pseudo_valid : pseudo_test) = (1,296 : 1,363) tasks

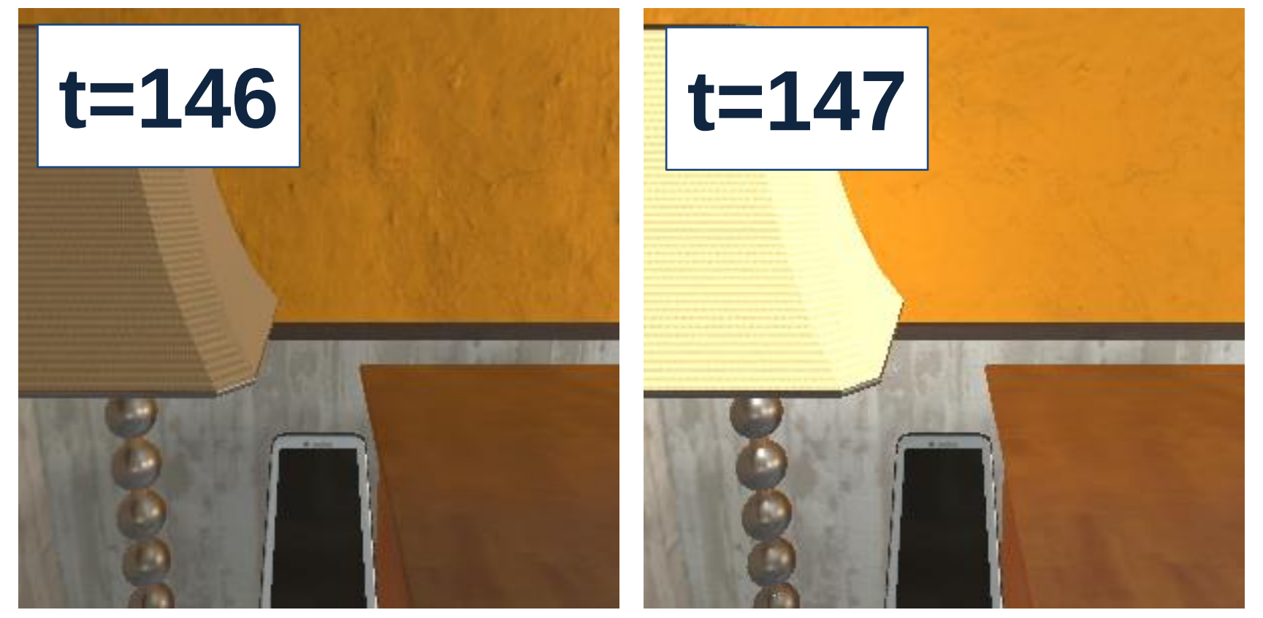
Method	Pseudo_Test SR↑	Pseudo_Test PWSR↑	Test SR↑
Baseline [Gao+, RA-L22]	0.31	0.19	-
Ours (w/o MAT)	0.34	0.20	-
Ours (w/ CLIP text encoder)	0.35	0.22	-
Ours (MAT for action)	0.36	0.21	-
Ours (DialMAT)	**0.39**	**0.23**	**0.14**

Instruction: “Move to the desk”



☺ Navigate to the specified desk

Instruction: “Move to the floorlamp power on the floorlamp”



☺ Navigate to the appropriate location and executed the appropriate action

Conclusions

- Introduced MAT to incorporate adversarial perturbations into the latent spaces of language, image, and action
- Introduced a crossmodal parallel feature extraction mechanisms to both language and image using foundation models

Our code available