

Alignment with the Dynamics of Visual Environments

(d) Bird's Eye View

Instruction: put the gold vase into safe.

Agent's action: Target object:

✔ Action succeeded
⚠ Action failed due to hallucination
⊘ Action failed due to invalid grammar

(a) GPT-4V

(b) InstructBLIP

(c) Our EMMA

Imitation Learning between a VLM student and an LLM teacher

Visual State

Task

Inference

Training

Retrospective Process

• Cross modality imitation in every interaction step via **DAgger + Direct Preference Optimization**

$$\mathcal{L}_{\text{imit}}(\cdot) \triangleq \log \sigma \left(\beta \log \frac{\pi_{\theta}(x_a^* | s_v)}{\pi_{\text{ref}}(x_a^* | s_v)} - \beta \log \frac{\pi_{\theta}(x_a | s_v)}{\pi_{\text{ref}}(x_a | s_v)} \right)$$

• The student's action is **negative**.
• The teacher's action is **positive**.

New SOTA and Qualitative Results on Visual ALFWorld Testing Environments

A. Template Task Instruction

Agent	Visual Env.	Textual Env.	Avg.
VLMs	ResNet-18* [56]	✗	0.06(-)
	MCNN-FPN* [56]	✗	0.05(-)
LLMs	BUTLER* [56]	✗	0.26(-)
	GPT-BUTLER [41]	✗	0.69(18.8)
	ReAct [73]	✗	0.54(20.5)
	Reflexion [54]	✗	0.91(18.7)
	DEPS* [61]	✗	0.76(-)
	AutoGen* [64]	✗	0.77(-)
VLMs	MiniGPT-4 [26]	✗	0.16(26.9)
	BLIP-2 [34]	✗	0.04(29.5)
VLMs	LLaMA-Adapter [17]	✗	0.13(27.5)
	InstructBLIP [11]	✗	0.22(26.2)
	EMMA (Ours)	✗	0.82(19.5)
	Human Performance* [55]	✓	0.91(-)

Cross-modality Imitation Learning

B. Free-form Task Instruction

Agent	Visual Env.	Textual Env.	Avg.
LLMs	BUTLER* [56]	✗	0.03(-)
	GPT-BUTLER [41]	✗	0.31(24.7)
	ReAct [73]	✗	0.37(23.6)
	Reflexion [54]	✗	0.78(17.0)
VLMs	MiniGPT-4 [76]	✓	0.00(30.0)
	BLIP-2 [34]	✓	0.01(29.7)
	LLaMA-Adapter [17]	✓	0.02(29.6)
	InstructBLIP [11]	✓	0.01(29.8)
	EMMA (Ours)	✓	0.68(22.0)

Distribution of Verbs for Human-annotated Instructions

Distribution of Verbs for Templated Instructions

- Grounding vision-language models (VLMs) pretrained on static image-text pairs as multi-model agents necessitate the alignment with the dynamics of visual environments.
- LLMs from *Parallel TextWorld* can provide accurate feedback to finetune VLMs for aligning with dynamic environments.

VLM Agent in Ai2Thor Visual World

The action taken by VLM agent

LLM Agent in TextWorld

```

Welcome to the TextWorld!
Your task is to: put a ...
...
You arrive at loc 2. On the table 1, you see knife 1, apple 1, cup 1...
> take apple 1 from table 1
You pick up the apple 1 from the table 1.
...
You arrive at loc 4. On the basin 1, you see ...
> clean apple 1 with basin 1
You clean the apple 1 using the basin 1.
...
You open the fridge 1. The fridge 1 is open. In it, you see egg 1, cup 2 ...
> put apple 1 in/on fridge 1
                    
```

The action taken by LLM agent

Target object