EMMA: Embodied Multi-Modal Agent trained by an LLM from a Parallel TextWorld

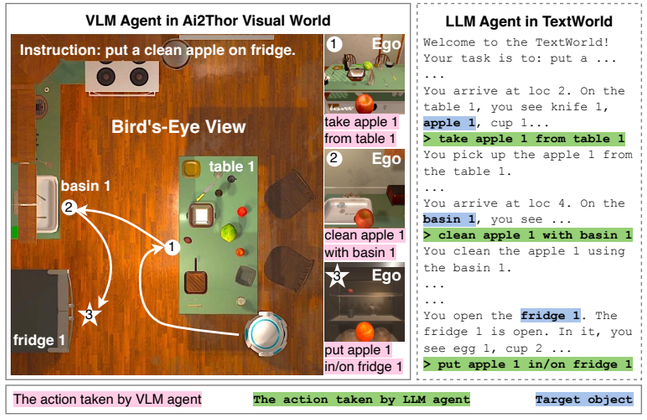
Yijun Yang[1,5,4], Tianyi Zhou[2], Kanxue Li[3], Dapeng Tao[3], Lusong Li[4], Li Shen[4,*], Xiaodong He[4], Jing Jiang[5], Yuhui Shi[1,*]

[1]Southern University of Science and Technology, [2]University of Maryland College Park, [3]Yunnan University,
[4]JD Explore Academy, [5]University of Technology Sydney

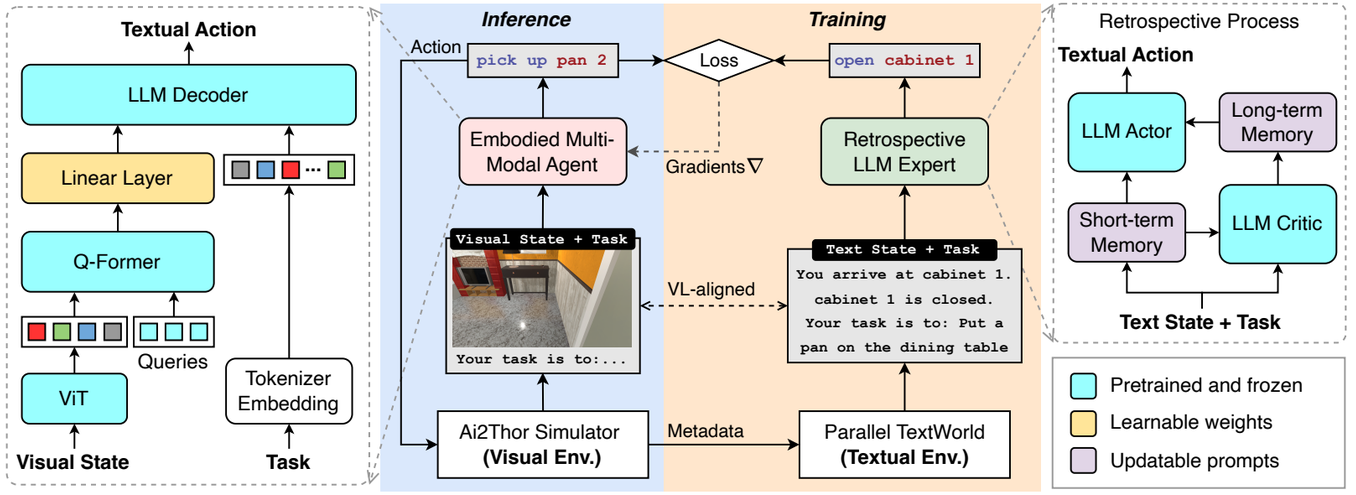
GitHub Link

Alignment with the Dynamics of Visual Environments

- 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.



Imitation Learning between a VLM student and an LLM teacher



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

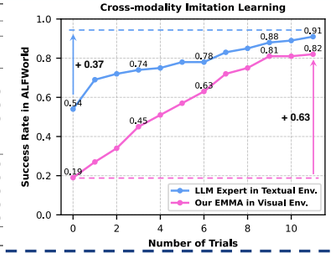
$$\mathcal{L}_{\text{imit}}(\cdot) \triangleq \log \sigma\left(\beta \log \frac{\pi_\theta(x_a^*|s_v)}{\pi_{\text{ref}}(x_a^*|s_v)} - \beta \log \frac{\pi_\theta(x_a|s_v)}{\pi_{\text{ref}}(x_a|s_v)}\right)$$

- 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.
VMs	ResNet-18* [56]	✓	✗	0.06 (-)
	MCNN-FPN* [56]	✓	✗	0.05 (-)
LMs	BUTLER* [56]	✗	✓	0.26 (-)
	GPT-BUTLER [41]	✗	✓	0.69 (18.8)
	ReAct [73]	✗	✓	0.54 (20.6)
	Reflexion [54]	✗	✓	**0.91** (18.7)
	DEPS* [61]	✗	✓	0.76 (-)
	AutoGen* [64]	✗	✓	0.77 (-)
VLMs	MiniGPT-4 [76]	✓	✗	0.16 (26.9)
	BLIP-2 [34]	✓	✗	0.04 (29.5)
	LLaMA-Adapter [17]	✓	✗	0.13 (27.5)
	InstructBLIP [11]	✓	✗	0.22 (26.2)
	EMMA (Ours)	✓	✗	0.82 (19.5)
	Human Performance* [55]	✓	✗	**0.91** (-)



B. Free-form Task Instruction

	Agent	Visual Env.	Textual Env.	Avg.
LMs	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)