From Observation to Abstractions: Efficient In-Context Learning from Human Feedback and Visual Demonstrations for VLM Agents

Gabriel Sarch[1] **Lawrence Jang**[1] **Michael J. Tarr**[1]

William W. Cohen[1,2] **Kenneth Marino**[2] **Katerina Fragkiadaki**[1]

[1]Carnegie Mellon University [2]Google DeepMind

https://ical-learning.github.io

Abstract

We propose an efficient method, In-Context Abstraction Learning (ICAL), to improve in-context VLM agents from sub-optimal demonstrations and human feedback. Specifically, given a noisy demonstration for a task in a new domain, LLMs/VLMs are used to fix inefficient actions and annotate four types of cognitive abstractions. These abstractions are then refined by executing the trajectory in the environment, guided by natural language feedback from humans. We demonstrate that this method rapidly learns useful experience abstractions. Our ICAL agent improves on the state-of-the-art when tested in dialogue-based instruction following in household environments in TEACh, action anticipation in Ego4D, and in multimodal autonomous web agents in VisualWebArena. In TEACh, we improve on the state-of-the-art by 12.6% in goal-condition success, outperforming LLM agents that use the raw visual demonstrations as in context examples without abstraction learning. In VisualWebArena, we improve on the state-of-the-art by an absolute 8.4% and relative 58.74% in task success, outperforming VLM agents that use hand-written examples. In Ego4D, we improve 6.4 noun and 1.7 action edit distance over few-shot GPT4V. Lastly, we find that weight fine-tuning and in-context abstraction learning complement each other, with their combination yielding the best performance.

1 Introduction

Humans acquire skills through language and observation, a model for automated systems. These systems must learn from verbal instructions and demonstrations to develop rapid learning technologies. This involves integrating linguistic feedback and demonstrative learning to refine knowledge across different contexts.

Research has used large language models (LLMs) and visual language models (VLMs) to derive insights from experiences, improving performance by adding these insights to prompts [6, 10, 7, 8]. However, there remains limitations in task transfer and underutilization of visual data.

We introduce a new method, In-Context Abstraction Learning (ICAL), for teaching VLMs using sub-optimal demonstrations and feedback. ICAL helps VLMs create and refine multimodal abstractions, aiding them in understanding task dynamics and critical knowledge [11, 1, 2, 4].

2 In-Context Abstraction Learning (ICAL)

ICAL starts by obtaining a noisy trajectory. It has two phases: (1) the abstraction phase $F_{abstract}$, where a VLM corrects the trajectory and adds language comments in isolation (Section 2.1), and (2)

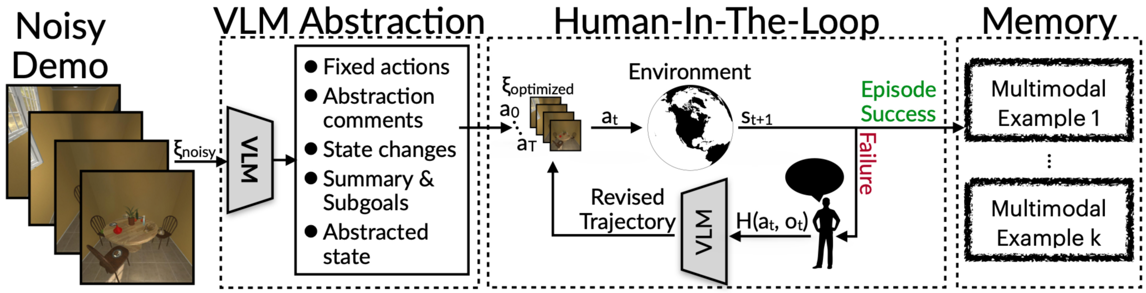

Figure 1: ICAL transforms raw experience into useful abstractions for in-context learning. *Top:* Given a noisy trajectory, It prompts a VLM to optimize actions and add language annotations. The optimized trajectory is executed, incorporating human feedback on failures. Successful examples are stored for future VLM in-context action generation.

the human-in-the-loop phase F_{hitl}, where the trajectory is executed with human feedback to refine it (Section 2.2). Each corrected trajectory is stored as a contextual reference for learning and inference.

2.1 VLM-driven Abstraction Generation

The abstraction function $F_{abstract}$ processes trajectory ξ_{noisy} into an optimized sequence ξ_{opt} with language abstractions L based on the instruction I and previous successful examples $\{e^1, \ldots, e^k\}$: $F_{abstract} : (\xi_{noisy}, I, \{e^1, \ldots, e^k\}) \rightarrow (\xi_{opt}, L)$. Given a noisy demonstration of actions and images, the VLM is prompted to annotate subgoals [2], causal relationships [11], state changes [1], and relevant state [4], highlighting important demonstration aspects.

2.2 Human-in-the-loop Abstraction Verification

Human-in-the-loop learning involves executing the optimized trajectory ξ_{opt} in the environment. A human monitors and provides feedback $H(a_t, o_t)$ on failures. The VLM is then prompted to revise the trajectory and annotations: $\Xi_{update}(\xi_{opt}, H(a_t, o_t), L, I, \{e^1, ..., e^k\}) \rightarrow \xi'_{opt}, L'$. The environment is reset after feedback, and the process repeats until the task is successful or a limit is reached.

2.3 Agent Deployment After Example Learning

Once examples are learned, the agent uses them to perform new tasks. The VLM generates actions based on the new instruction I, visual and textual state, and retrieving the top K ICAL examples from the learned set E to guide action generation, with similarity scores based on input instruction, textual, and visual state features. Implementation uses `gpt-4-1106-vision-preview` for the text generation, unless otherwise noted.

3 Experiments

TEACh Evaluation ICAL shows continual improvement in validation task success as more examples are learned. ICAL outperforms baseline approaches significantly, achieving a 17.9% absolute improvement in task success rate over unprocessed demonstrations. ICAL outperforms the handwritten examples used by the existing state-of-the-art in TEACh by 12.6% in goal condition success and 0.6% in task success (Table 2).

Improving with Fine-Tuning Fine-tuning the LLM on ICAL examples further improves performance, especially when combined with retrieval-augmented generation, indicating the utility of integrating learned examples in training (Table 5).

Ablation Studies Ablation studies confirm that each component of ICAL—from the abstraction phase to the human-in-the-loop phase—is crucial for achieving the observed improvements in performance (Table 2).

Ego4D Evaluation See Table 3. ICAL demonstrates superior few-shot performance on Ego4D action anticipation compared to hand-written few-shot GPT4V that uses chain of thought [13] by 6.4 noun and 1.7 action edit distance. ICAL also remains competitive with the fully supervised baseline [3] despite using 639x less training data. We find GPT4V video processing to have the most trouble with verb prediction.

Visual Web Navigation Evaluation ICAL demonstrates state-of-the-art performance on VisualWebArena, outperforming previous best methods by an absolute 8.4% (relative 58.74%) in success rate (Table 4).

Figure 2: Evaluation of the TEACh unseen validation set using GPT-3.5-1106. Visual demos utilize an inverse dynamics model, while Kinesthetic demos are labeled with ground truth actions. GC = goal-condition success

	Success	GC
HELPER handwritten [9]	34.5	36.7
Zero-Shot CoT [6]	11.8	24.6
Raw Visual Demos	17.2	26.6
Raw Kinesthetic Demos	26.5	29.5
ICAL (ours)	**35.1**	**49.3**
w/o abstraction phase	29.4	44.9
w/o human-in-the-loop	29.9	41.0
w/ re-ranking [12]	**35.3**	**51.7**
w/ GPT4	41.7	63.6

Figure 3: Evaluation on Ego4D Long Term Action Anticipation unseen validation subset. ICAL does not use human-in-the-loop due to the passive nature of this task.

	ED@(Z=20)		
	Verb	Noun	Action
Supervised [3] 639x more data	0.725	0.739	0.923
Few-shot CoT [13]	0.787	0.757	0.941
ICAL (ours)	**0.780**	**0.693**	**0.924**

Figure 4: Evaluation results on VisualWebArena. Ablations are done on a reduced 257 episodes.

	Seen	Unseen	Avg.
GPT4V+SoM [5]	16.3	14.1	14.3
ICAL (ours)	**38.8**	**20.9**	**22.7**
Ablations			
GPT4V+SoM [5]	11.5	12.9	12.7
ICAL (ours)	28.0	**21.6**	22.2
w/o image	28.0	17.3	19.0
w/ full trajectory	**57.7**	**21.6**	**25.5**

Figure 5: Results on finetuning the LLM on the ICAL demonstrations. The model used is GPT3.5-turbo-1106.

	Success	GC
zero-shot	11.8	24.6
retrieval	35.1	49.3
finetuned	23.2	40.3
finetuned + retrieval	35.8	54.2

References

[1] Lisa Feldman Barrett and Moshe Bar. See it with feeling: affective predictions during object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521): 1325–1334, 2009.

[2] Matthew M Botvinick, Yael Niv, and Andew G Barto. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *cognition*, 113(3):262–280, 2009.

[3] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

[4] Mark K Ho, David Abel, Carlos G Correa, Michael L Littman, Jonathan D Cohen, and Thomas L Griffiths. People construct simplified mental representations to plan. *Nature*, 606(7912):129–136, 2022.

[5] Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*, 2024.

[6] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

[7] Bodhisattwa Prasad Majumder, Bhavana Dalvi Mishra, Peter Jansen, Oyvind Tafjord, Niket Tandon, Li Zhang, Burch Callison-Burch, and Peter Clark. Clin: A continually learning language agent for rapid task adaptation and generalization. *arXiv*, 2023.

[8] Jesse Mu, Victor Zhong, Roberta Raileanu, Minqi Jiang, Noah Goodman, Tim Rocktäschel, and Edward Grefenstette. Improving intrinsic exploration with language abstractions (2022). *URL https://arxiv. org/abs/2202.08938*.

[9] Gabriel Sarch, Yue Wu, Michael Tarr, and Katerina Fragkiadaki. Open-ended instructable embodied agents with memory-augmented large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

[10] Noah Shinn, Beck Labash, and Ashwin Gopinath. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*, 2023.

[11] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

[12] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[13] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.