Knowing More, Acting Better: Hierarchical Representation for Embodied Decision-Making

Chunhui Zhang[1] **Zhongyu Ouyang**[1] **Xingjian Diao**[1]
Zheyuan Liu[2] **Soroush Vosoughi**[1]
[1]Department of Computer Science, Dartmouth College
[2]Department of Computer Science and Engineering, University of Notre Dame
{chunhui.zhang.gr, zhongyu.ouyang.gr, xingjian.diao.gr, soroush.vosoughi}@dartmouth.edu
zliu29@nd.edu

Abstract

Modern embodied AI uses multimodal large language models (MLLMs) as policy models, predicting actions from final-layer hidden states. This widely adopted approach, however, assumes that monolithic last-layer representations suffice for decision-making—a structural simplification *at odds with* decades of cognitive science, which highlights the importance of distributed, hierarchical processing for perception and action. Addressing this foundational asymmetry, we introduce a hierarchical action probing method that explicitly aggregates representations from all layers, mirroring the brain's multi-level organization. Experiments reveal that early layers facilitate spatial grounding, middle layers support contextual integration, and later layers enable abstract generalization—which shows MLLMs inherently encode distributed action-relevant structures. These layer-wise features are integrated by a lightweight probe for spatial reasoning and contextual understanding, without costly backbone fine-tuning. This hierarchical solution shows significant improvements over standard last-layer embodied models in physical simulators, achieving a 46.6% success rate and a 62.5% gain in spatial reasoning tasks. These findings challenge conventional assumptions in embodied AI, establishing hierarchical probing as a principled alternative grounded in both cognitive theory and empirical evidence.

1 Introduction

In embodied AI, multimodal large language models (MLLMs) enable agents to perceive visual-language inputs and generate actionable outputs (Driess et al., 2023; Zitkovich et al., 2023; Szot et al., 2024). Their current application in embodied tasks predominantly relies on single-layer representations—specifically, the final hidden state—for action prediction, mirroring traditional language generative modeling.

However, this reliance on monolithic, final-layer representations presents a *structural asymmetry* in embodied AI: this largely unchallenged convention contrasts sharply with decades of cognitive science emphasizing **hierarchical, distributed perception** (Paccanaro and Hinton, 2001; Hinton, 2023). This simplification, while perhaps adequate for static language tasks, overlooks rich intermediate-layer features in embodied multimodal tasks. Our work challenges the common assumption that final-layer outputs are sufficient for complex embodied policies. We hypothesize that MLLM layers inherently specialize (e.g., early for spatial relations, middle for contextual dependencies, later for abstract semantics), necessitating stratified grounding. Discarding this internal structure potentially creates a bottleneck in aligning with the cognitive complexity of embodied environments. Indeed, even in pure language tasks, leveraging multiple layers can improve performance (Abbas et al., 2024; Chételat et al., 2025; Tian et al., 2025), yet many embodied models (Zitkovich et al., 2023) still predominantly use final-layer outputs.

To explore this, we introduce Hierarchical Action Probing (HAP) as a *conceptual pivot*. To the best of our knowledge, this is the first attempt to employ **probing techniques**—originally for model interpretability—to uncover the phenomenon of how distributed hierarchical representations, which are fundamental to human cognition (Hasson et al., 2008; Lerner et al., 2011), also naturally emerge and can be harnessed in embodied policy models. In particular, HAP recenters the internal layer-wise states of MLLM-based policy models for grounding behavior, using the full representational capacity. Rather than simply combining multiple layers as in prior NLP work, *our novel contribution* lies in improving *embodied policy models through cognitive insights*. We show that the hierarchical distributed representation is crucial in both human cognition and multimodal embodied tasks.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19144–19155
November 4-9, 2025

As a direct consequence of this insight, our technical approach consists of (i) a visual encoder, (ii) the MLLM backbone generating layer-wise states, and (iii) a multi-layer action probe *hierarchically aggregating* these states. This hierarchy allows policy models to capture fine-grained spatial relationships and high-level task semantics, mirroring the brain's distributed processing. The methodological improvements are valuable outcomes stemming from this cognitive perspective. We examine HAP's effectiveness on language-guided rearrangement tasks in the physical simulators.

Experiments show HAP achieves strong performance, with our 7B model matching LLaRP-7B's 42% success rate and our 13B model reaching 46.6%, surpassing LLaRP-13B's 46%. HAP particularly excels in spatial reasoning (13% vs. 8% for LLaRP-7B). Detailed ablation studies further reveal layer-wise specializations—intermediate layers for spatial understanding, later layers for object affordances—empirically supporting our hierarchical design principles and the benefits of probing these internal structures.

2 Related Work

Recent research has explored using large pretrained MLLMs (Jian et al., 2023, 2024; Han et al., 2024; Zhang et al., 2025; Diao et al., 2025; You et al., 2025; Zhou et al., 2025; Guo et al., 2025; Wang et al., 2025a,b; Liu et al., 2023, 2024) as policies for embodied tasks, often termed vision-language-action (VLA) models. Approaches like RT-2-X (Zitkovich et al., 2023) and OpenVLA (Kim et al., 2024) scale to billion-parameter policies, leveraging large datasets like Open X-Embodiment for training. Other works, such as (Huang et al., 2024; Li et al., 2024; Zhen et al., 2024; Dorka et al., 2024), focus on single-robot or simulated setups, while (Zitkovich et al., 2023; O'Neill et al., 2024) lack efficient fine-tuning for new configurations. Parameter-efficient techniques, such as LLaRP (Szot et al., 2024) and OpenVLA (Kim et al., 2024), address computational challenges by fine-tuning MLLMs, enabling generalization to novel tasks.

In contrast to existing VLA models that rely on monolithic last-layer representations, we show the role of hierarchical representations in embodied policy models—mirroring the human brain's hierarchical and distributed cognitive processes.

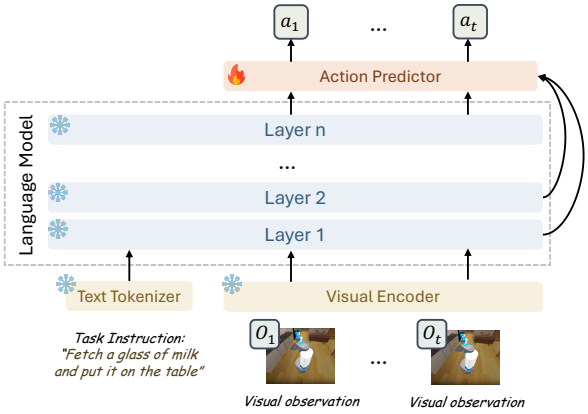

Figure 1: We probe hidden states across all LM layers to predict actions through an action probing module.

3 Methodology

The embodied task is formulated as a Partially Observable Markov Decision Process (POMDP), defined by $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \rho_0, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{O}$ the observation space (e.g., egocentric RGB images), $\mathcal{A}$ the action space, $\mathcal{P}$ transition dynamics, $\mathcal{R}$ the reward function, ρ_0 the initial state distribution, and γ the discount factor. The agent, given a natural language goal $g \in \mathcal{G}$, generate actions $a_t \in \mathcal{A}$ based on visual observations $o_t \in \mathcal{O}$ to maximize cumulative reward $\mathcal{R}(s_t, g)$. The objective is to learn a *goal-conditioned policy* $\pi(a_t|o_1, \ldots, o_t, g)$. While POMDPs formally use belief states, conditioning on history $o_1, \ldots, o_t$ is a practical approach with sequence models like LMs, which implicitly summarize history to approximate the belief state, addressing partial observability. This policy model can generalize to novel goal distributions $\mathcal{G}'$.

3.1 Architecture of the Policy Model

Our architecture (Fig. 1) comprises three key components: the visual encoder, the LM backbone, and the multi-layer action probe.

Visual Encoder The visual encoder $E^V_\phi : \mathcal{O} \to \mathbb{R}^D$ (a pretrained ViT with a learnable MLP) processes raw visual observations o_t into visual token embeddings $v_t = E^V_\phi(o_t)$. This maps high-dimensional visual inputs to a compact token space for the MLLM backbone.

LM as Policy The LM backbone ψ_θ, with L transformer layers, processes concatenated visual token embeddings v_t and language goal embeddings $E^T_\theta(g)$. For each layer $l \in \{1, \ldots, L\}$ and time t, it outputs hidden states

	Overview			Train and New Scenes		Paraphrastic Robustness					Behavior Generalization			
	All	**BG**	**PR**	**TR**	**SC**	**INS**	**RE**	**SP**	**CTX**	**IR**	**MR**	**NO**	**MO**	**CO**
ZS-ChatGPT	22.0	23.0	21.0	57.0	52.0	58.0	24.0	5.0	10.0	11.0	24.0	61.0	2.0	5.0
ZS-Llama	12.0	14.0	10.0	54.0	41.0	34.0	3.0	0.0	5.0	6.0	6.0	50.0	0.0	0.0
ZS-Flamingo	6.0	8.0	4.0	24.0	14.0	18.0	0.0	0.0	0.0	2.0	8.0	24.0	2.0	0.0
LSTM-Flan	25.0 ± 1	28.0 ± 1	23.0 ± 1	98.0 ± 1	95.0 ± 8	85.0 ± 2	6.0 ± 1	4.0 ± 4	15.0 ± 3	5.0 ± 2	19.0 ± 4	83.0 ± 3	0.0 ± 0	10.0 ± 6
LSTM-Llama	2.0 ± 1	0.0 ± 0	3.0 ± 2	31.0 ± 2	15.0 ± 2	12.0 ± 3	1.0 ± 2	2.0 ± 4	0.0 ± 1	0.0 ± 1	0.0 ± 1	1.0 ± 1	0.0 ± 0	0.0 ± 0
LLaRP-scratch	17.0 ± 4	18.0 ± 5	16.0 ± 3	90.0 ± 9	90.0 ± 9	59.0 ± 13	3.0 ± 1	3.0 ± 3	4.0 ± 3	13.0 ± 4	15.0 ± 6	58.0 ± 16	0.0 ± 0	1.0 ± 1
LLaRP-7B	42.0 ± 2	45.0 ± 3	38.0 ± 1	**99.0 ± 1**	96.0 ± 4	92.0 ± 2	26.0 ± 2	8.0 ± 1	34.0 ± 2	32.0 ± 2	47.0 ± 5	95.0 ± 4	0.0 ± 1	39.0 ± 3
Ours-7B	42.4	43.1	41.0	**99.0**	97.4	91.3	28.2	13.6	37.9	35.1	48.5	89.2	0.3	36.4
LLaRP-13B	46.0	48.0	44.0	**98.0**	**100.0**	**95.0**	31.0	15.0	41.0	37.0	51.0	**98.0**	0.0	**45.0**
Ours-13B	**46.6**	**48.3**	**45.0**	**99.0**	97.1	94.6	**33.4**	**18.7**	**42.1**	**38.0**	**53.5**	95.4	**2.1**	43.6
harder setting:														
LSTM-Flan	12.0	14.0	11.0	57.0	52.0	50.0	3.0	0.0	0.0	2.0	11.0	43.0	0.0	0.0
LLaRP-7B	**28.0**	**27.0**	28.0	56.0	61.0	**62.0**	23.0	1.0	32.0	**24.0**	**31.0**	**56.0**	0.0	20.0
Ours-7B	26.4	24.9	**29.4**	**60.2**	**63.1**	59.7	**26.0**	**4.6**	**35.8**	23.9	27.1	48.6	0.0	**21.7**

Table 1: Success rate across tasks for all baselines and settings. BG stands for behavior generalization, PR for paraphrastic robustness, TR for train, SC for new scenes, INS for instruction rephrasing, RE for referring expressions, SP for spatial relationships, CTX for context, IR for irrelevant text, MR for multiple rearrangements, NO for novel objects, MO for multiple objects, and CO for conditional instructions. App. B.1 reports results on Qwen2.5 (Qwen Team, 2025), which further suggest that HAP is effective for other recent LMs.

$h_t^l \in \mathbb{R}^D$. The first layer's state is $h_t^1 = \psi^1\left(\text{Concat}(E_\theta^T(g), v_1, \ldots, v_t)\right)$. For $l > 1$, hidden states are $h_t^l = \psi^l(h_t^{l-1})$. This recursive expression is a notational simplification; the frozen LM backbone inherently includes standard components like pre-normalization and residual connections, which remain unmodified by our approach. Our shallow probe operates on these frozen representations, avoiding deep gradient flow issues. This hierarchical processing allows the LM to capture diverse features for embodied grounding.

Multi-Layer Action Probe The action probe aggregates hidden states $h_t^1, \ldots, h_t^L$ from all layers corresponding to visual tokens (language tokens are excluded). Each h_t^l is reduced to 512 dimensions via a 2-layer MLP (ReLU, LayerNorm). These are concatenated and passed to a final 2-layer MLP $P(\cdot)$ for action prediction:

$$\pi(a_t|o_1, \ldots, o_t, g) = P\left(\text{Concat}\left(\text{MLP}_1(h_t^1), \ldots, \text{MLP}_L(h_t^L)\right)\right). \quad (1)$$

It provides disentangled, parallel access to representations across the hierarchy, allowing each layer's output to contribute more independently to the prediction. This contrasts with standard residual connections where earlier layer information flows indirectly and can become entangled or attenuated. Our approach allows learnable weighting of layer-wise specializations (e.g., spatial, contextual, abstract), enhancing performance, particularly in tasks requiring nuanced spatial reasoning, by leveraging this richer, more accessible hierarchical information. For discussion of how the probing's disentangled layer access contrasts with residual connections, see App A.4. Ablation studies of the probing usage are in App. B and Tab. 2.

4 Experiments

We assess our approach on challenging language-guided rearrangement tasks within the Habitat 2.0 simulator (Szot et al., 2021). For *fair comparisons*, we follow the usage of frozen VisualCortex ViT (Majumdar et al., 2023) and Llama-2 (Touvron et al., 2023)/Qwen2.5-VL (Qwen Team, 2025) backbones from Szot et al. (2024), training only a lightweight hierarchical probe using PPO; baselines include: 1) ZS-ChatGPT, iteratively refining actions with environment feedback; 2) ZS-LLaMA, generating a single plan with LLM (Touvron et al., 2023); 3) ZS-Flamingo, a multimodal zero-shot planner using IDEFICS (Alayrac et al., 2022); 4) LLaRP-Scratch, a 7B-LLM variant trained from scratch; 5) LSTM-Flan, combining Flan-T5 (Chung et al., 2024) with an LSTM; and 6) LSTM-LLaMA, using LLaMA with a Perceiver Resampler (Jaegle et al., 2022). Further details on simulators, baselines, and setups are in App. A.

4.1 Main Results

According to Tab. 1, our hierarchical probing shows strong performance across the benchmark suite, with Ours-7B achieving a 42% overall success rate (matching LLaRP-7B) and Ours-13B reaching 46.6% (surpassing LLaRP-13B's 46%). These results significantly outperform zero-shot baselines like ChatGPT (22%), LLaMA (12%), and Flamingo (6%), highlighting the effectiveness of unlocking pretrained capabilities by post-training.

Our usage of hierarchical representation

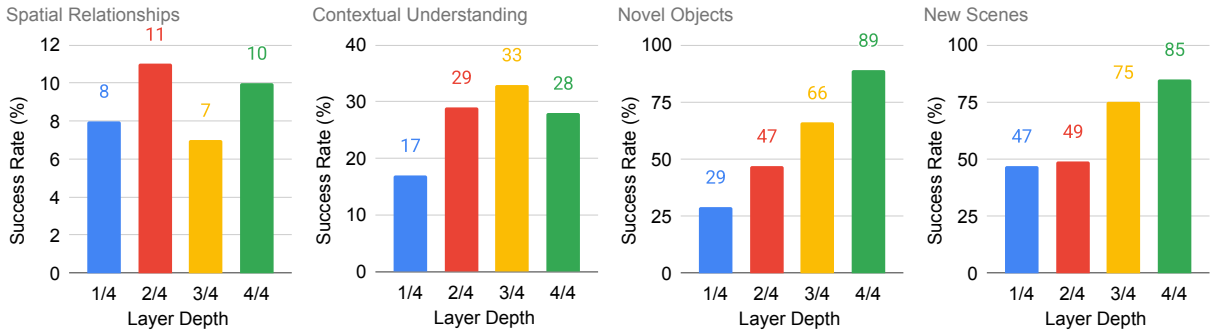

Figure 2: Layer-wise ablation study across tasks show distinct specialization patterns: early layers (1-8) excel at spatial reasoning, middle layers (9-24) optimize contextual understanding, and later layers (25-32) enable abstract reasoning for tasks like novel objects and new scenes.

shows particular strength in tasks requiring complex spatial and contextual reasoning. Specifically, Ours-7B achieves a 62.5% relative improvement in spatial relationship tasks (13% vs. 8% for LLaRP-7B) and an 8% improvement in paraphrastic robustness (41% vs. 38%). The larger Ours-13B model further extends these gains, demonstrating improvements in contextual understanding (42% vs. 41% for LLaRP-13B) and multiple rearrangements (53% vs. 51%). These results validate our cognitive hypothesis that integrating features across different processing levels enhances the model's ability to handle complex spatial-temporal relationships.

While this approach performs well at spatial-temporal tasks, we observe some limitations in tasks heavily dependent on high-level semantic understanding. Notably, performance slightly decreases in novel object handling (89% v.s. 95% for LLaRP-7B) and remains unchanged for multiple object manipulation (0% for both models). This pattern suggests that our current implementation, while effective at integrating diverse features, may partially dilute the rich semantic representations typically concentrated in the LM's final layers.

To assess real-world applicability, we also evaluate our method under more stringent conditions where invalid actions result in immediate episode termination and explicit stop actions are required. In this challenging setting, Ours-7B demonstrates improved robustness, achieving better performance in spatial reasoning (4% vs. 1% for LLaRP-7B) and paraphrastic understanding (29.4% vs. 28%). These results indicate that our hierarchical approach enhances the model's precision and reliability in more realistic scenarios.

4.2 LM Hierarchies Mirror Neural Processing

To understand the hierarchical representations, in Fig. 2, we conduct layer-wise ablation studies by dividing the LM's 32 layers into quarters.

Early and Middle Layers (1-16) These layers predominantly process spatial and geometric information. Mid-early layers (8-16) achieve peak performance in spatial reasoning (11% success rate vs. 8% for layers 1-8), suggesting their importance for processing low-level visual features and spatial relationships. Similarly, contextual understanding improves progressively through middle layers, reaching 33% success rate at layers 16-24.

Later Layers (17-32) These layers specialize in abstract reasoning and generalization. Performance on novel object tasks increases dramatically from 29% (early layers) to 89% (full model), while scene adaptation improves from 47% to 85%. However, we observe a slight decline in contextual understanding (28% vs. 33%) when including the final layers, indicating a trade-off between abstraction and contextual processing.

These findings show the complementary nature of layer-wise representations: early layers process spatial information of the received observation visual tokens, middle layers integrate context, and later layers enable abstraction. Additionally, App. D has qualitative behavioral examples showing how HAP improves decision-making in tasks requiring spatial and contextual reasoning.

5 Conclusion

Inspired by cognitive science, this work challenges prior MLLM's use in embodied AI. Our hierarchical action probing, by engaging internal MLLM structures beyond just final layers, improves pol-

icy models—notably in spatial reasoning and contextual understanding. We also reveal MLLMs' inherent, human brain-like layer-wise specializations, affirming distributed hierarchical processing as pivotal for more capable embodied agents.

Limitations

While our approach performs well on many tasks, it is slightly less effective when it requires only high-level semantic understanding (such as novel object generalization). Future work could focus on better balancing hierarchical integration and semantic preservation. Additionally, as MLLMs are developed toward more flexible action spaces and deployed in the real world, greater effort will be needed to ensure they remain safe and harmless to human society, and to reduce potential risks.

Acknowledgment

This research was supported in part by the National Science Foundation under Grant No. 2452367.

References

Momin Abbas, Yi Zhou, Parikshit Ram, Nathalie Baracaldo, Horst Samulowitz, Theodoros Salonidis, and Tianyi Chen. 2024. Enhancing in-context learning via linear probe calibration. In *International Conference on Artificial Intelligence and Statistics*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: part 3.1, knowledge storage and extraction. In *Proceedings of the 41st International Conference on Machine Learning*.

Didier Chételat, Joseph Cotnareanu, Rylee Thompson, Yingxue Zhang, and Mark Coates. 2025. Innerthoughts: Disentangling representations and predictions in large language models. In *International Conference on Artificial Intelligence and Statistics*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*.

Xingjian Diao, Chunhui Zhang, Weiyi Wu, Zhongyu Ouyang, Peijun Qing, Ming Cheng, Soroush Vosoughi, and Jiang Gui. 2025. Temporal working memory: Query-guided segment refinement for enhanced multimodal understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

Nicolai Dorka, Chenguang Huang, Tim Welschehold, and Wolfram Burgard. 2024. What matters in employing vision language models for tokenizing actions in robot control? In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: an embodied multimodal language model. In *International Conference on Machine Learning*.

Siddhartha Gairola, Moritz Böhle, Francesco Locatello, and Bernt Schiele. 2025. How to probe: Simple yet effective techniques for improving post-hoc explanations. In *The Thirteenth International Conference on Learning Representations*.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Jiawei Guo, Feifei Zhai, Pu Jian, Qianrun Wei, and Yu Zhou. 2025. Crop: Contextual region-oriented visual token pruning. *arXiv preprint arXiv:2505.21233*.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*.

Xiaotian Han, Yiren Jian, Xuefeng Hu, Haogeng Liu, Yiqi Wang, Qihang Fan, Yuang Ai, Huaibo Huang, Ran He, Zhenheng Yang, and Quanzeng You. 2024. InfiMM-webmath-40b: Advancing multimodal pre-training for enhanced mathematical reasoning. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Uri Hasson, Eunice Yang, Ignacio Vallines, David J Heeger, and Nava Rubin. 2008. A hierarchy of temporal receptive windows in human cortex. *Journal of neuroscience*.

Geoffrey Hinton. 2023. How to represent part-whole hierarchies in a neural network. *Neural Computation*.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. 2024. An embodied generalist agent in 3d world. In *International Conference on Machine Learning*.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2022. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2023. Bootstrapping vision-language learning with decoupled language pre-training. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yiren Jian, Tingkai Liu, Yunzhe Tao, Chunhui Zhang, Soroush Vosoughi, and Hongxia Yang. 2024. Expedited training of visual conditioned language generation via redundancy reduction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. 2024. OpenVLA: An open-source vision-language-action model. In *Annual Conference on Robot Learning*.

Yulia Lerner, Christopher J Honey, Lauren J Silbert, and Uri Hasson. 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of neuroscience*.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. 2024. Vision-language foundation models as effective robot imitators. In *International Conference on Learning Representations*.

Yanqing Liu, Xianhang Li, Zeyu Wang, Bingchen Zhao, and Cihang Xie. 2024. Clips: An enhanced clip framework for learning with synthetic captions. *arXiv preprint arXiv:2411.16828*.

Yanqing Liu, Kai Wang, Wenqi Shao, Ping Luo, Yu Qiao, Mike Zheng Shou, Kaipeng Zhang, and Yang You. 2023. Mllms-augmented visual-language representation learning. *arXiv preprint arXiv:2311.18765*.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. 2023. Where are we in the search for an artificial visual cortex for embodied intelligence? In *Thirty-seventh Conference on Neural Information Processing Systems*.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *IEEE International Conference on Robotics and Automation*.

Alberto Paccanaro and Geoffrey E Hinton. 2001. Learning hierarchical structures with linear relational embedding. In *Advances in neural information processing systems*.

Alibaba Group Qwen Team. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Dachuan Shi, Yonggan Fu, Xiangchi Yuan, Zhongzhi Yu, Haoran You, Sixu Li, Xin Dong, Jan Kautz, Pavlo Molchanov, et al. 2025. Lacache: Ladder-shaped kv caching for efficient long-context modeling of large language models. *International Conference on Machine Learning*.

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimír Vondruš, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. 2021. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems*.

Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazoure, Rin Metcalf, Walter Talbott, Natalie Mackraz, R Devon Hjelm, and Alexander T Toshev. 2024. Large language models as generalizable policies for embodied tasks. In *International Conference on Learning Representations*.

Yijun Tian, Xingjian Diao, Ming Cheng, Chunhui Zhang, Jiang Gui, Soroush Vosoughi, Xiangliang Zhang, Nitesh V Chawla, and Shichao Pei. 2025. On the design choices of next level llms. *Authorea Preprints*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.

Junxi Wang, Jize Liu, Na Zhang, and Yaxiong Wang. 2025a. Consistency-aware fake videos detection on short video platforms. In *International Conference on Intelligent Computing*. Springer.

Junxi Wang, Yaxiong Wang, Lechao Cheng, and Zhun Zhong. 2025b. Fakesv-vlm: Taming vlm for detecting fake short-video news via progressive mixture-of-experts adapter. *arXiv preprint arXiv:2508.19639*.

Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. 2020. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *International Conference on Learning Representations*.

Zhifan Ye, Kejing Xia, Yonggan Fu, Xin Dong, Jihoon Hong, Xiangchi Yuan, Shizhe Diao, Jan Kautz, Pavlo Molchanov, and Yingyan Celine Lin. 2025. Longmamba: Enhancing mamba's long context capabilities via training-free receptive field enlargement. *arXiv preprint arXiv:2504.16053*.

Wenhao You, Xingjian Diao, Chunhui Zhang, Keyi Kong, Weiyi Wu, Zhongyu Ouyang, Chiyu Ma, Tingxuan Wu, Noah Wei, Zong Ke, et al. 2025. Music's multimodal complexity in avqa: Why we need more than general multimodal llms. *arXiv preprint arXiv:2505.20638*.

Xiangchi Yuan, Yijun Tian, Chunhui Zhang, Yanfang Ye, Nitesh V Chawla, and Chuxu Zhang. 2024. Graph cross supervised learning via generalized knowledge. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Xiangchi Yuan, Chunhui Zhang, Zheyuan Liu, Dachuan Shi, Soroush Vosoughi, and Wenke Lee. 2025. Superficial self-improved reasoners benefit from model merging. *arXiv preprint arXiv:2503.02103*.

Chunhui Zhang, Yiren Jian, Zhongyu Ouyang, and Soroush Vosoughi. 2025. Pretrained image-text models are secretly video captioners. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 2024. 3d-vla: a 3d vision-language-action generative world model. In *International Conference on Machine Learning*.

Changshi Zhou, Haichuan Xu, Ningquan Gu, Zhipeng Wang, Bin Cheng, Pengpeng Zhang, Yanchao Dong, Mitsuhiro Hayashibe, Yanmin Zhou, and Bin He. 2025. Language-guided long horizon manipulation with llm-based planning and visual perception. *arXiv preprint arXiv:2509.02324*.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*.

A Detailed Experimental Setting

To ensure clarity and reproducibility, this section explicitly details the experimental framework, baselines, and model setup adopted in our study. Our methodology, including task descriptions, datasets, and evaluation metrics, aligns with those established in the LLaRP paper (Szot et al., 2024), ensuring comparability and consistency with existing literature in language-driven embodied AI.

A.1 Task Setting

We evaluate our models within the domain of language-driven rearrangement tasks. In these tasks, embodied agents operate in simulated environments and are instructed via natural language to manipulate and reposition objects. The tasks probe a range of capabilities, including:

- **Spatial Relationships:** Requiring rearrangement based on nuanced spatial descriptions (e.g., "to the right of the left counter").
- **Novel Objects:** Handling objects not seen during training.
- **Contextual Reasoning:** Acting on implied rather than explicit cues.
- **Multiple Rearrangements:** Following multi-step or composite instructions.
- **Referring Expressions:** Grounding indirect object descriptions (e.g., "a yellow curved fruit").
- **Conditional Instructions:** Executing conditional logic in instructions (e.g., "If the fridge is open, move X; otherwise, move Y").

A.2 Dataset and Simulator

Experiments utilize the **Habitat 2.0 simulator environment** (Szot et al., 2021), specifically its language-guided rearrangement benchmark. This provides diverse scenarios categorized by varying complexity, generalization requirements (e.g., new scenes, new objects), and linguistic challenges, forming a robust testbed for embodied agent generalization.

A.3 Baselines

Following LLaRP (Szot et al., 2024), we include a range of baselines. Of particular note, the "LLaRP-Scratch" model (Szot et al., 2024) uses the same architecture as LLaMA-7B but is trained *from scratch* without any LLM pretraining (approx. 2B parameters). This baseline, which we include for context rather than direct comparison, isolates the impact of LLM pretraining: as shown in prior work, pretrained LLaRP models converge faster and generalize better than their scratch-trained counterparts, despite having more parameters. This comparison reinforces the motivation for our approach, which leverages frozen, pretrained LLMs for improved generalization in embodied tasks.

Other baselines include:

- **ZS-ChatGPT** uses ChatGPT to iteratively refine a high-level action plan with proprioceptive feedback.
- **ZS-LLaMA** uses instruction-tuned LLaMA-65B (Touvron et al., 2023) to generate a static action plan.
- **ZS-Flamingo** leverages IDEFICS, a Flamingo-based (Alayrac et al., 2022) VLM, for single-shot planning.
- **LSTM-Flan** encodes instructions with Flan-T5 (Chung et al., 2024) and observations with an LSTM.
- **LSTM-LLaMA** encodes instructions with LLaMA and observations with a Perceiver Resampler (Jaegle et al., 2022) plus LSTM.

A.4 Model Setup

All baselines and our proposed model employ the VisualCortex model (Majumdar et al., 2023), a ViT backbone optimized for egocentric visual tasks. In our approach, rather than training a new end-to-end vision-language model, we combine a *frozen* VisualCortex ViT visual encoder with a *frozen* LLaMA-2 backbone. These two modules are connected via a lightweight trainable MLP that serves as a modality connector. The output of the ViT is treated as a sequence of additional visual tokens, appended after the language instruction tokens. The entire concatenated sequence is then jointly processed by the LLaMA-2 backbone.

At each timestep t, tokenized language instructions g and visual observations $o_1, \ldots, o_t$ are embedded and passed through the frozen LLaMA-2 backbone, which has L transformer layers. The l-th layer outputs a hidden state h_t^l. Our HAP module then aggregates the hidden states from all layers $(h_t^1, \ldots, h_t^L)$ that correspond to the visual tokens

(importantly, hidden states corresponding to language instruction tokens are excluded at this stage). Each layer's visual hidden state h_t^l is first projected into a 512-dimensional space using a 2-layer MLP (with ReLU activations and LayerNorm). These projected vectors are then concatenated and passed through a final 2-layer MLP to predict a distribution over actions a_t, as formally defined in Equation 1 in the main paper.

Distinction from Standard Residual Connections A critical aspect of our HAP design is its difference from how information is typically propagated via standard residual connections in transformer architectures like LLaMA-2. Mainstream generative decoder LLMs employ residual connections, often with pre-normalization, allowing each layer to pass its representation forward, e.g., $h_t^l = \text{block}^l(h_t^{l-1}) + h_t^{l-1}$, followed by $h_t^{l+1} = \text{block}^{l+1}(h_t^l) + h_t^l$. While this formulation facilitates information flow through the network, it does so *indirectly*. Each layer primarily accesses the output of its immediate predecessor, and information from earlier layers (e.g., h_t^{l-3}) becomes absorbed and transformed through successive operations. Consequently, the contribution from these earlier layers can become increasingly entangled and potentially attenuated as network depth increases. The final layer's output, therefore, contains a blended representation of all preceding layers.

In stark contrast, our HAP module *explicitly extracts* the hidden state h_t^l from *each* transformer layer independently. These layer-specific representations are then fed in parallel into our lightweight probing module. This architectural choice provides **disentangled, parallel access** to representations across the entire hierarchy of the LLM. Each layer's output can thus contribute more independently to the final action prediction, rather than its influence being solely mediated through subsequent layers and the final output layer, as the final output in our study is $y_t = \sum_{l=1}^{L} \text{block}^l\left(h_t^{l-1}\right) + x_t$.

The benefits of this disentangled access are manifold. As our layer-wise ablation studies demonstrate (Fig. 2 in the main paper), different layers specialize in processing distinct aspects of the input: early layers are more attuned to spatial reasoning, middle layers to contextual understanding, and later layers to abstract generalization. HAP's ability to independently tap into these specialized representations allows for a more nuanced and dynamic integration of information. This hierarchical integration enables our model to outperform last-layer baselines, particularly in tasks demanding robust spatial reasoning (e.g., achieving a +62.5% relative gain over LLaRP-7B in SP tasks, as shown in Table 1). In summary, while residual connections ensure latent information flows through the network, they do not offer the same level of independent access or the capacity for dynamic, task-adaptive weighting of layer-specific features that our probing approach enables.

Action Prediction from Visual Tokens A key design choice within HAP is that, for the final action prediction, our output module (D_ω, embodied by the final MLPs in the probe) operates *only* on the hidden states corresponding to the visual observation tokens $(o_1, \ldots, o_t)$ from all transformer layers. Hidden states corresponding to the language instruction tokens are excluded from this final aggregation step. The motivation behind this is to ensure that the predicted action a_t is fundamentally a reaction to the agent's current perceived state, as processed and contextualized by the LLM in light of the overall language goal g. Although only the visual hidden states are directly fed into the action output module, the LLM's self-attention mechanism ensures that these visual hidden states are already deeply contextually enriched by the preceding language instruction tokens. Therefore, the action output module D_ω effectively learns a mapping from a goal-conditioned, multi-layer visual state representation to an appropriate action distribution. This approach, adapted from and extending LLaRP (Szot et al., 2024), allows the policy to harness features across different levels of abstraction, improving generalization across both spatial and semantic tasks.

Online Experience Collection for PPO During the PPO training phase, the "dataset" is not a static collection but rather an online stream of experience. This experience is collected as the agent interacts with the language-guided rearrangement tasks within the Habitat 2.0 simulator (Szot et al., 2021). Each collected episode comprises a natural language instruction, a (potentially novel) visual scene, and sparse rewards indicating task progress or completion. The agent selects high-level actions from a predefined skill set (e.g., pick, place, navigate), and its policy is optimized via PPO using these collected trajectories of observations, actions, and rewards.

A.5 Training Protocol

Training begins with policy pretraining using last-layer hidden states and a basic action decoder, implemented via the *DD-PPO algorithm* (Wijmans et al., 2020), a multi-GPU adaptation of PPO. Next, we fine-tune the multi-layer probing module while freezing the MLLM backbone and visual encoder. This two-stage protocol preserves pretrained representations and mitigates catastrophic forgetting(Alayrac et al., 2022; Yuan et al., 2025). Only the probing module and action prediction head are optimized for cumulative reward using DD-PPO updates, reducing computational overhead by avoiding backward passes through the full MLLM. Training hyperparameters include a discount factor $\gamma = 0.99$, Generalized Advantage Estimation (GAE) with $\lambda = 0.95$, learning rate of $1e{-}4$, and batch size 64. Training runs for 70 hours on 4 NVIDIA H100 GPUs.

A.6 Evaluation Metrics

The primary evaluation metric is the **Success Rate**—the percentage of episodes in which the agent completes the given rearrangement task according to the language instruction. Following LLaRP (Szot et al., 2024), we report overall and fine-grained success rates, analyzing generalization to new scenes, paraphrastic robustness, and behavioral generalization (e.g., handling semantically similar but syntactically different tasks).

By following these established settings and explicitly detailing the role of hierarchical probing over visual token representations, our work aims to contribute directly comparable and interpretable findings to the embodied AI research landscape.

B Rationale for Multi-Layer Probe Architecture

In designing the multi-layer action probe, careful consideration was given to the architecture of the MLP used to process hidden states from each transformer layer. We select a 2-layer MLP based on a balance between representational capacity and training complexity, a choice informed by established practices in recent probing literature (Gairola et al., 2025; Gao et al., 2025). To validate this choice, we conduct experiments comparing probes of varying depths.

As shown in Table 2, a 2-layer MLP achieved the optimal trade-off in performance. The decision for a 2-layer MLP is further supported by the following observations:

Method	Total (%)	PR (%)	BG (%)
LLaRP-13B	46.0	44.0	48.0
· 1-layer MLP	42.0	39.9	44.0
· **2-layer MLP (*ours*)**	**46.6**	**45.0**	**48.3**
· 3-layer MLP	42.5	40.1	45.0

Table 2: Comparison of action probe MLP depths. LLaRP refers to the baseline performance using its standard configuration. PR and BG mean tasks in Paraphrastic Robustness and tasks in Behavior Generalization. Our 2-layer MLP probe demonstrates superior or comparable performance to other configurations.

First, a single-layer (linear) probe proved insufficient for capturing the complex semantic-to-action mappings inherent in the middle-layer representations of pretrained MLLMs. These mappings frequently necessitate non-linear transformations, which a linear probe cannot adequately model, thereby leading to degraded performance.

Second, conversely, deeper probes (e.g., a 3-layer MLP) introduce increased training complexity and a higher risk of overfitting to the training data. Given that the primary objective of probing is to assess the expressiveness of existing pretrained representations rather than to learn entirely new ones, a more complex and deeper architecture could diminish stability without yielding substantial performance improvements.

Therefore, the 2-layer MLP configuration offers an optimal balance, effectively capturing necessary non-linearities while minimizing training overhead and overfitting risks. This aligns with conventions in recent studies on model interpretability and probing (Gairola et al., 2025; Gao et al., 2025; Yuan et al., 2024).

B.1 Generalizability to Other Model Architectures

To address the generalizability of HAP beyond the LLaMA-2 architecture, we initiate experiments with the Qwen series models. While these investigations are ongoing, early results are promising and provide initial insights into HAP's adaptability. The experimental settings are kept consistent with our default setup.

Preliminary findings, presented in Table 3, indicate that HAP can be effectively applied to models like Qwen2.5-7B (with a VC-1 ViT visual encoder), which exhibits solid performance. We hypothesize that its larger pretraining corpus (18T tokens com-

Method	Total (%)	BG (%)	PR (%)
Qwen2.5-VL-7B	29.5	35.2	23.8
Qwen2.5-7B w/ VC-1 ViT	37.5	40.0	35.2
Ours-7B (LLaMA-2)	42.4	43.1	41.0
Ours-7B (Q2.5-VL-7B w/ VC-1)	48.6	50.4	46.8

Table 3: Preliminary generalizability results with current Qwen series models. PR and BG mean tasks in Paraphrastic Robustness and tasks in Behavior Generalization. Ours-7B refers to our HAP method applied to LLaMA-2 7B for baseline comparison. The latest Qwen2.5-7B with VC-1 ViT visual encoder yields the best results.

pared to LLaMA-2's 2T tokens) contributes to this outcome, suggesting that HAP can leverage the strong pretrained capabilities of diverse foundation models. Conversely, Qwen2.5-VL-7B, when using its native built-in visual encoder, showed less effective performance. This is attributed to the specific nature of embodied tasks, where egocentric video data plays a crucial role in adaptation—a factor potentially less emphasized in models pretrained on more generic vision datasets. When we use the latest Qwen2.5-VL-7B with the VC-1 ViT visual encoder, it achieves the best results so far. This suggests our approach can keep improving with stronger foundation models.

Despite these initial variations, we are optimistic about the broader applicability of our method, particularly when paired with foundation models with different architectures (Gu and Dao, 2023; Ye et al., 2025; Shi et al., 2025) that possess strong pretrained representations relevant to embodied AI.

C Ablation Study on Skip-Layer Partitioning

Our main paper presents an ablation study where transformer layers were partitioned into consecutive blocks. To further investigate the impact of layer selection, we conducted an additional ablation study employing a skip-layer strategy, where representations are probed from every other layer. Our initial intent was to establish a baseline with a general, consecutive partitioning strategy.

The results of this skip-layer probing are presented in Table 4. These findings suggest that sparse probing (e.g., using a skip-layer strategy) can be as effective as, and in some aspects slightly outperform, dense probing of all consecutive layers. For many subtasks, performance remained comparable or showed slight improvements. This phenomenon, where sparse probing yields strong

Method	Total (%)	BG (%)	PR (%)
LLaRP-7B	42.0	45.0	38.0
Ours-7B (skip one)	42.2	45.0	39.4
Ours-7B (default all)	42.4	43.1	41.0

Table 4: Ablation study results with skip-layer probing strategy. "LLaRP" is the baseline. "Ours (skip one)" refers to probing every other layer. "Ours (default all)" refers to our standard HAP method probing all layers.

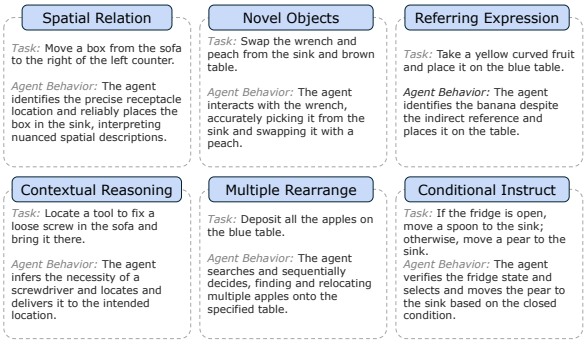

Figure 3: Examples of agent behaviors across diverse and challenging scenarios: (a) Spatial Relationships, where location descriptions are correctly interpreted; (b) Novel Objects, demonstrating interaction with previously unseen items; (c) Referring Expressions, involving identification of objects from indirect descriptions; (d) Contextual Reasoning, where implicit needs are inferred and acted upon; (e.e.) Multiple Rearrangements, showcasing handling of multi-object tasks; and (f) Conditional Instructions, requiring adherence to logical conditions.

results, aligns with observations in prior works on interpretability and sparse probing techniques (Gurnee et al., 2023; Allen-Zhu and Li, 2024).

D Behavior Study: Qualitative Analysis of HAP Capabilities

To illustrate how HAP enhances decision-making, Fig. 3 presents successful agent behaviors in challenging scenarios, highlighting benefits of leveraging multi-layer MLLM representations.

Spatial Reasoning (Fig. 3a): The agent correctly interprets complex spatial instructions (e.g., relocating a box to a nuanced target like "right of the left counter," identified as the sink). HAP strengthens this by utilizing specialized lower-layer MLLM information attuned to geometric details, increasing accuracy and stability in such spatial tasks.

Novel Objects (Fig. 3b): The agent's interaction with previously unseen items, like swapping a novel wrench, showcases HAP's potential for adaptability. By enabling a more balanced seman-

tic integration from various MLLM layers, HAP may improve the recognition and handling of new objects, though careful design is needed to preserve high-level semantics.

Referring Expressions (Fig. 3c): Successful grounding of indirect descriptions (e.g., identifying a "yellow curved fruit" as a banana) is demonstrated. HAP reinforces this by better integrating visual features from early MLLM layers with semantic cues from middle and later layers, leading to more robust understanding of such language.

Contextual Reasoning (Fig. 3d): HAP-enabled agents demonstrate effective inference from context, such as finding a screwdriver for a loose screw without explicit mention. HAP's leveraging of middle MLLM layers for contextual integration enhances the understanding of implicit cues, improving performance in these context-driven tasks.

Multiple Rearrangements (Fig. 3e): HAP shows strength in handling tasks involving multiple objects, like depositing all apples on a designated table. Its hierarchical integration improves task robustness, ensuring comprehensive completion even when the number of items is not explicitly enumerated.

Conditional Instructions (Fig. 3f): The agent correctly executes conditional logic, for instance, moving a pear because a fridge was observed to be closed. HAP supports stronger and more explicit conditional reasoning by effectively integrating information needed to evaluate conditions and select appropriate subsequent actions, thereby improving decision consistency over simpler models.

E Robustness of the Improvements

To assess the robustness of our improvements, we conduct paired *t*-tests and Wilcoxon signed-rank tests on representative tasks, using results from three independent experimental runs per setting in Table 5.

Task	Baseline	Ours	Paired *t*-test (p)	Wilcoxon (p)
SR	14.5, 15.2, 16.0	19.3, 18.1, 18.9	0.031	0.250
CR	39.8, 41.1, 41.7	41.5, 43.2, 41.5	0.533	0.500

Table 5: Statistical significance analysis of representative tasks. Results are averaged over three runs. SR means Spatial Relationships. CR means Contextual Reasoning.

For **Spatial Relationships**, the improvement is statistically significant at the 5% level using the paired *t*-test ($p = 0.031$), indicating a reliable benefit of hierarchical aggregation in this domain. For **Contextual Reasoning**, the observed gain is not statistically significant, which we attribute to the higher run-to-run variability characteristic of embodied environments. These findings support our central claim: hierarchical aggregation provides consistent and statistically significant gains in tasks benefiting from enriched grounding, while improvements in other domains may be more modest or variable due to randomness introduced by simulators.