From Perception to Action: Building Spatially Intelligent Embodied Agents

Abstract

Building artificial agents in complex physical environments requires moving beyond passive vision-language understanding toward active, tool-augmented, and self-reflective multimodal systems. While recent foundation models have demonstrated remarkable capabilities in multimodal understanding, they often lack the spatial awareness, procedural reasoning, and adaptive decision-making needed for real-world interaction. In this work, I propose to advance Vision-Language-Action (VLA) foundation models through post-training frameworks that integrate spatial intelligence, tool-augmented reasoning, and self-reflection mechanisms. My recent studies like *Video-STAR* and *AutoDrive-R^2* demonstrate that reinforcing multimodal agents with explicit reasoning loops and external tool interfaces significantly enhances their zero-shot generalization and decision reliability in dynamic environments like video understanding and autonomous driving. These efforts form the foundation of my postdoctoral research agenda, which aims to develop interpretable, socially aligned VLA systems capable of truthful reasoning and adaptive tool use across domains ranging from robotics to scientific discovery.

Keywords: Vision-Language Models, Vision-Language Action, Embodied Agents, Spatial Intelligence

1 Introduction

Recent advances in Vision-Language Models (VLMs) have enabled unprecedented capabilities in multimodal understanding. Vision-Language Models (VLMs) such as CLIP [1], Flamingo [2], and LLaVA [3] have demonstrated strong zero-shot transfer across diverse perception tasks, while emerging Vision-Language-Action (VLA) models like RT-2 [4] and OpenVLA [5] aim to bridge perception with motor control for embodied agents. These developments suggest a promising path toward general-purpose AI that can "see, reason, and act" in real-world environments.

Despite these progress, current foundation models still struggle to reason robustly and adaptively in dynamic 3D physical world. This limitation mainly stems from three aspects: ❶ **Cross-modal hallucinations:** Existing approaches rely heavily on text-based chain-of-thought (CoT) reasoning for visual understanding, which leads to misalignment between temporal visual features with discrete textual representations. ❸ **Missing Self-reflection Capabilities:** Foundation models lacks rarely exhibit self-awareness about their own uncertainty or errors. Such absence may cause model overconfidently execute unsafe maneuvers, or fail silently when faced with distribution shifts. ❷ **Lacking Spatial Intelligence:** Most VLMs simply treat inputs as collections of 2D image patches and fail to establish a grounded mapping from 2D visual pixels to 3D geometry, fundamentally limiting their capacity for safe embodiment. Consequently, current VLA methods exhibit hallucinated actions or unsafe decisions when applied to high-stakes domains like scientific automation.

To address these, my recent research tackles each of these limitations through targeted innovations in multimodal post-training: Firstly, to mitigate cross-modal hallucinations, we introduced *Video-STAR* [6], a framework that equips VLMs with external perception tools including visual detector, pose estimator, and online retrieval-augmentation (RAG) during inference. By grounding textual reasoning in explicit, temporally consistent visual signals, Video-STAR enables robust zero-shot video understanding while effectively enhancing cross-modal interleaving through multimodal CoT reasoning.

Second, to address the lack of self-reflection capabilities, we proposed *AutoDrive-R^2* [7], a metacognitive architecture for VLA models in autonomous driving. In this framework, the agent iteratively critiques its own action plans using environmental feedback and revises its decisions through a reflection loop. This mechanism not only improves driving safety and reliability but also provides a generalizable template for self-correcting behavior in embodied agents.

Thirdly, recognizing that both tool-augmentation and self-reflection require a foundation of spatial intelligence, my approach is deeply informed by my earlier work on 3D scene reconstruction. Specifically, I developed series of geometrically grounded multi-view stereo methods including *SD-MVS* [8], *DVP-MVS* [9], and *MSP-MVS* [10] to recover coherent 3D structures from real-world imagery across diverse global scenarios. This line of work honed my expertise in modeling metric geometry, occlusion reasoning, and cross-modal alignment, which are essential capabilities for establishing the pixel-to-world mapping missing in current VLMs. It also shaped my core perspective: embodied intelligence must be treated not merely as a language or policy optimization problem, but as a physically grounded reasoning challenge.

Building on this triad of insights, my postdoctoral research will integrate these threads into a unified framework for trustworthy VLA systems. Specifically, I will advance: ❶ **Tool-augmented inference** to align multimodal reasoning with perceptual reality and suppress hallucinations; ❷ **Self-reflective policy learning** to enable agents to monitor, critique, and correct their own behavior; and ❸ **Spatially grounded pretraining** thats leverage 3D geometric priors to instill intrinsic spatial intelligence into foundation models from the outset. This integrated approach will bridge the gap between passive multimodal understanding and active, reliable agency. By jointly optimizing for perceptual fidelity, metacognitive control, and physical grounding, the proposed research will deliver both theoretical advances in embodied AI and practical systems for high-stakes applications in autonomous driving, robotic automation, and AI for scientific discovery. Ultimately, it aims to move beyond "smart perception" toward responsible, capable, and truly embodied intelligence.

2 Proposed Research Plan

2.1 Unified Multimodal Encoding via Tool-Augmented Representation Learning

I will develop a *unified multimodal encoding framework* in which external perception tools (e.g., pose estimators, video clipper, online RAG) are not merely invoked at inference time, but actively shape the internal representation space of the VLA model during training. Rather than treating tools as external plugins invoked during decoding, I will design a transformer architecture with *modality-agnostic tokenization* that natively embeds structured tool outputs (e.g., bounding boxes, depth maps, trajectory proposals) alongside image patches and text tokens. This enables joint attention across perception, language, and action primitives, allowing the model to reason holistically. For example, aligning "grasp the red cup" with precise 3D coordinates and gripper kinematics in a single forward pass. The goal is a foundation model that inherently understands *what can be done* as part of *what is seen and said*.

2.2 Embodied Intelligence with Long-Horizon Planning & Continual Skill Learning

I will advance self-reflective policy learning beyond single-step correction toward *long-horizon embodied reasoning* and *lifelong skill acquisition*. Specifically, I propose to develop hierarchical reflection mechanisms that decompose complex tasks (e.g., "prepare a lab experiment") into subgoals, critique plan feasibility at multiple time scales, and dynamically replan when failures occur. Crucially, I will investigate *continual learning in embodied agents*: as the agent acquires new skills (e.g., using a new instrument), how can it retain and compose prior knowledge without catastrophic forgetting? I will explore memory-augmented architectures and skill-factorized policy representations that enable safe, compositional transfer, thereby turning self-reflection into a driver of scalable, adaptive intelligence in open-world environments.

2.3 Spatially Grounded Pretraining for Intrinsic 3D Reasoning

I will develop spatially grounded foundation models that directly map visual inputs to *metrically accurate action commands*, closing the loop between scene understanding and physical interaction. Instead of outputting abstract actions like "move forward," the model will predict *quantitative spatial transformations*, e.g., "rotate gripper by 15°," "translate arm by 0.3m along the surface normal", conditioned on fine-grained 3D scene geometry. This requires learning a *visuo-motor spatial encoder* that translates pixel observations into a coordinate-aware action space, leveraging geometric priors (e.g., depth, surface normals, affordance maps) during pretraining. The vision is to build agents that "understand space like humans do": judging distances, anticipating occlusions, and reasoning about object affordances to generate safe, precise, and context-aware actions—turning spatial intelligence into embodied competence.

3 Broader Impacts

This research advances the development of *human-centered AI design*: interpretable planning, continual learning without catastrophic forgetting, and metrically accurate actions all contribute to systems that are not only more effective but also more predictable and controllable by human users. By endowing agents with spatial intelligence, self-reflection, and grounded action generation, the work directly addresses safety and reliability concerns in high-stakes domains such as autonomous transportation, assistive robotics for the elderly, AI for science, and AI-driven scientific automation.

4 Alignment with Lab

My research on spatially intelligent embodied agents closely aligns with Prof. Kai-Wei Chang's expertise in trustworthy multimodal foundation models and constrained reasoning. Her pioneering work on VisualBERT, GLIP, and DesCo demonstrates a strong commitment to grounding language in visual perception—complementing my focus on 3D spatial grounding and tool-augmented reasoning in VLA systems. Moreover, her lab's emphasis on aligning AI with human values, robustness, and constraint-aware reasoning directly supports my goals of building self-reflective, hallucination-resistant agents that act safely in real-world environments. This synergy offers a fertile ground for advancing trustworthy, embodied AI.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[4] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[5] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[6] Zhenlong Yuan, Xiangyan Qu, Chengxuan Qian, Rui Chen, Jing Tang, Lei Sun, Xiangxiang Chu, Dapeng Zhang, Yiwei Wang, Yujun Cai, and Shuo Li. Video-star: Reinforcing open-vocabulary action recognition with tools. *arXiv preprint arXiv:2510.08480*, 2025.

[7] Zhenlong Yuan, Jing Tang, Jinguo Luo, Rui Chen, Chengxuan Qian, Lei Sun, Xiangxiang Chu, Yujun Cai, Dapeng Zhang, and Shuo Li. Autodrive-r2: Incentivizing reasoning and self-reflection capacity for vla model in autonomous driving. *arXiv preprint arXiv:2509.01944*, 2025.

[8] Zhenlong Yuan, Jiakai Cao, Zhaoxin Li, Hao Jiang, and Zhaoqi Wang. Sd-mvs: Segmentation-driven deformation multi-view stereo with spherical refinement and em optimization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 6871–6880, 2024.

[9] Zhenlong Yuan, Jinguo Luo, Fei Shen, Zhaoxin Li, Cong Liu, Tianlu Mao, and Zhaoqi Wang. Dvp-mvs: Synergize depth-edge and visibility prior for multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9743–9752, 2025.

[10] Zhenlong Yuan, Cong Liu, Fei Shen, Zhaoxin Li, Jinguo Luo, Tianlu Mao, and Zhaoqi Wang. Msp-mvs: Multi-granularity segmentation prior guided multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9753–9762, 2025.