From Representation to Reasoning: Toward General-Purpose Visual Intelligence

Chen Wei

Rice University
chen.wei@rice.edu

Abstract

This talk surveys my research agenda on advancing general-purpose visual intelligence, moving AI beyond static recognition toward active reasoning and embodied action. A central challenge is enabling AI systems to generalize reliably in low-data and long-tail regimes. I address this by combining multimodal representation learning with agentic reasoning frameworks such as PyVision, which equips vision models to dynamically generate tools for deliberate problem-solving, and ViGaL, which leverages gameplay to instill transferable cognitive skills for reasoning under scarcity. These efforts chart a trajectory from representation and generation to interactive, embodied agents, re-imagining AI as an active collaborator capable of tool use, imagination, and purposeful engagement across both digital and physical environments.

My research seeks to advance AI systems from passive perception to active reasoning and embodied intelligence. At the intersection of vision, language, and robotics, I develop algorithms that combine representation learning, generative modeling, and agentic reasoning, discussed as follow.

Generalization in Low-Data and Long-Tail Scenarios. Despite impressive progress in high-resource settings, current AI systems are surprisingly brittle when data is scarce or imbalanced. My approach couples representation learning with *explicit reasoning* to generalize beyond memorization: (i) align vision–language representations for cross-domain transfer, and (ii) endow agents with tool-use so they can actively observe, reason and act.

PyVision. PyVision (Zhao et al. 2025) introduces a multi-turn agent that *dynamically generates Python tools at test time*—without a predefined toolset—to solve complex visual queries in a thought–action–observation loop. The agent composes on-the-fly tools, for example, zoom and crop, segmentation/thresholding, OCR, counting, geometric solvers, temporal search, and chains them with language instructions to gather targeted evidence, produce intermediate visualizations, and verify partial hypotheses before committing to an answer. Because tools are created and executed as code rather than baked into a monolithic network, PyVision can plug in off-the-shelf modules, reuse external knowledge, and remain sample-efficient in low-data regimes (few-shot VQA, fine-grained counting and localization, rationale-grounded captioning), while yielding transparent traces that improve interpretability, debuggability and trust.

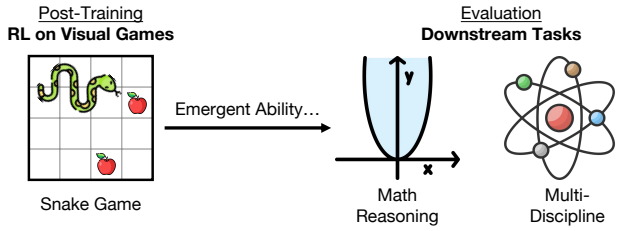

Figure 1: **Overview of ViGaL (Xie et al. 2025).** We propose a novel post-training paradigm where MLLMs are tuned via RL to play arcade-style games such as Snake. We demonstrate that gameplay post-training enables MLLMs to achieve out-of-domain generalization, enhancing their performance on downstream multimodal reasoning tasks requiring math, spatial and multi-discipline reasoning.

ViGaL (Play to Generalize). ViGaL (Xie et al. 2025) trains agents with reinforcement learning on *simple arcade-style games* to acquire reusable cognitive primitives—spatial working memory, causal prediction, and subgoal discovery—under controlled stochasticity and sparse reward. We show that policies or distilled representations learned through play transfer to a *suite of out-of-domain multimodal reasoning tasks*, for example, visual question answering and compositional puzzles, *without exposure to explicit reasoning supervision*, improving robustness in long-tail settings. The key idea is to replace scarce labeled data with structured *experience*: curriculum design, environment randomization, and temporal abstraction induce skills that later act as inductive biases for zero and few-shot generalization.

Taken together, PyVision and ViGaL point to a generalization strategy that pairs *evidence-seeking tool use* with *skill priors learned through interaction*, enabling models to adapt reliably in low-data and long-tail regimes.

Summary The talk will trace a trajectory from robust representation and generative modeling to structured reasoning and agentic behavior. The unifying theme is to reimagine AI not as a static predictor but as an active collaborator, bringing us closer to general-purpose intelligence.

References

Xie, Y.; Ma, Y.; Lan, S.; Yuille, A.; Xiao, J.; and Wei, C. 2025. Play to Generalize: Learning to Reason Through Game Play. *arXiv:2506.08011*.

Zhao, S.; Zhang, H.; Lin, S.; Li, M.; Wu, Q.; Zhang, K.; and Wei, C. 2025. PyVision: Agentic Vision with Dynamic Tooling. In *MTI-LLM @ NeurIPS*.