

Humans effortlessly navigate the complexities of the visual world, adapting to novel objects, scenes, and tasks with ease. After learning to place an apple on the left of a plate, we can easily adapt to placing it on the right. This flexibility stems from our ability to understand modular, high-level concepts (“*apple*”, “*plate*”, “*on-left*”) and ground them to low-level visual information (grounding “*on-left*” to relative spatial positions). Through multimodal interactions with the world, we develop this structured understanding that guides perception and reasoning across diverse tasks. In contrast, modern visual systems still struggle with this adaptability. They often fail to ground concepts to visual inputs, limiting their ability to generalize to novel reasoning tasks and tackle fundamental challenges like spatial reasoning.

Inspired by human cognition, I am pursuing a PhD to **develop machines with human-like visual intelligence—systems that learn and leverage a structured understanding of the visual world to reason as flexibly as humans**. I aim to explore questions such as:

- 1) What types of structure should models learn to enable generalization and adaptation across diverse visual reasoning tasks?
- 2) How can models learn representations of these structures from multimodal, in-the-wild data?
- 3) How can models leverage these structured representations to perform complex reasoning?
- 4) How can insights from human cognition guide the development of models to achieve such flexibility?

Over the past three years, I have led and contributed to several projects in the **Stanford Vision and Learning Lab** with **Prof. Fei-Fei Li, Prof. Jiajun Wu, and Prof. Tobias Gerstenberg**. My early research with Prof. Li sparked my interest in human-like visual reasoning and culminated in publications at **NeurIPS 2022, ICML 2023, and two NeurIPS 2023 Workshops (co-first author)** [1, 2, 3]. Subsequently, I led two projects [4, 5] that refined my interest in enabling machines to understand the structure of the visual world and developed the necessary technical, creative problem-solving, and interdisciplinary communication skills.

Learning Structured Representations for Visual Reasoning.

With Prof. Wu, I began exploring how to learn and leverage structured representations to improve generalization for downstream tasks, like state classification. This task involves determining whether a given state (“an *apple* is *next-to* a *plate*”) is True or False, supporting broader capabilities like scene understanding, robotic planning, and controlled generation. A key challenge is that real-world environments present countless variations in objects (“*apple*”, “*plate*”) and predicates (“*next-to*”), making it intractable to collect exhaustive training data. This necessitates models that generalize efficiently to novel states.

To address this, I proposed a method that leverages the semantic relationships between predicates to structure reasoning over scenes. Concretely, given an image and state, our method learns to map these inputs into a joint image-state representation space that captures the hierarchical relationships between predicates. For instance, while both “*next-to*” and “*on-left*” involve understanding distance, “*on-left*” requires more specific spatial reasoning. We infer the underlying predicate hierarchy using self-supervised contrastive losses, with pre-text tasks designed by an LLM to capture relationships between predicates. Additionally, we embed these representations in hyperbolic space, which has exponentially growing distances that enable our method to naturally encode the hierarchy.

Our method surpasses state-of-the-art supervised models, improving few-shot generalization to unseen states and transfer from sim-to-real images. It also outperforms inference-only VLMs, showcasing how structured representations can enable smaller models to achieve performance comparable to much larger models trained on vastly more data. Through driving this project from its initial proposal to submission at

ICLR 2025 (co-first author) [5], I learned to distill a broad challenge into a well-defined project and ground a high-level insight into a concrete, effective method.

Our recent submission to **CVPR 2025** [6] further emphasizes the value of structure in enabling efficient learning. We demonstrate that learning the structural components of real-world scenes (layouts, objects, and poses) improves the generation of realistic rooms from limited data.

Leveraging Insights from Cognition for Human-Like Reasoning.

To develop models with robust visual reasoning capabilities, I aim to draw inspiration from the cognitive processes humans use for reasoning. Previously, with Prof. Gerstenberg, I collaborated with an interdisciplinary team to explore: *How can insights into human causal inference enhance AI's reasoning about long sequences of events?* Motivated by how humans integrate visual, language, and auditory information to solve whodunit-style mysteries (“Who broke the plate?”), I spearheaded the development of **MARPLE**—a benchmark for long-horizon inference that challenges AI to identify which agent caused an environmental change using multimodal cues.

I developed the simulation infrastructure, designed human studies in collaboration with cognitive scientists, and proposed inference models inspired by human mental simulation with world models. Concretely, I formulated the inference problem as a Partially Observable Markov Decision Process (POMDP) and simulated agent trajectories via Monte Carlo rollouts with learned agent policies. These policies predict actions given a visual representation of the environment, requiring them to implicitly extract high-level information about agent behaviors from low-level visual inputs and reason effectively about them. Some policy variants also integrate language or audio cues to improve action predictions.

Despite substantial training, inference models struggled to generalize to new environments, and adding language and audio provided only modest improvements. In contrast, humans achieved superior performance and robustness, likely drawing from their structured understanding of the world to accurately synthesize and reason about multimodal inputs. We published these findings at the **NeurIPS 2024 Datasets and Benchmarks Track (co-first author)** [4], along with a complementary cognitive science study at **CogSci 2024** [7] which demonstrated how human-inspired multimodal simulations improved model alignment with human judgments. These experiences highlighted the gap between human and AI abilities to derive structure and reason based on low-level inputs, affirming my interest in bridging this gap through cognition-inspired, multimodal learning.

Future Work.

Building on my experience integrating cognitive insights and structured approaches, I aim to further explore how these methods can enhance AI’s ability to understand the structure of the visual world and reason flexibly based on this understanding. In particular, some directions that I hope to work on include:

- Understanding what kinds of structure emerges within general-purpose representations and their impact on downstream performance [8, 9]
- Learning visual concepts to understand the structure of the world and leveraging them across domains from visual reasoning to robotic planning [10-12]
- Endowing machines with human-like reasoning processes to further enable complex reasoning and adaptation to novel challenges, especially in visual domains with limited data [13-15]

Long-term, I aspire to become a professor, driving advancements toward human-like visual intelligence.

References

- [1] Luo, Zelun, Zane Durante*, Linden Li*, Wanze Xie, Ruochen Liu, **Emily Jin**, Zhuoyi Huang et al. "MOMA-LRG: Language-Refined Graphs for Multi-Object Multi-Actor Activity Parsing." *In NeurIPS Datasets and Benchmarks Track* (2022).
- [2] Kurenkov, Andrey, Michael Lingelbach, Tanmay Agarwal, **Emily Jin**, Chengshu Li, Ruohan Zhang, Li Fei-Fei, Jiajun Wu, Silvio Savarese, and Roberto Martín-Martín. "Modeling Dynamic Environments with Scene Graph Memory." *In ICML* (2023).
- [3] **Jin, Emily***, Jiaheng Hu*, Zhuoyi Huang, Ruohan Zhang, Jiajun Wu, Li Fei-Fei, and Roberto Martín-Martín. "Mini-BEHAVIOR: A Procedurally Generated Benchmark for Long-horizon Decision-Making in Embodied AI." Presented at the *NeurIPS Generalization in Planning Workshop* (2023) and the *NeurIPS Agent Learning in Open-Endedness Workshop* (2023).
- [4] **Jin, Emily***, Zhuoyi Huang*, Jan-Philipp Fränken, Weiyu Liu, Hannah Cha, Erik Brockbank, Sarah A. Wu, Ruohan Zhang, Jiajun Wu, and Tobias Gerstenberg. "MARPLE: A Benchmark for Long-Horizon Inference." *In NeurIPS Datasets and Benchmarks Track* (2024).
- [5] **Jin, Emily***, Joy Hsu*, Jiajun Wu. "Predicate Hierarchies Improve Few-Shot Classification." *Under Review at ICLR* (2025).
- [6] Hsu, Joy, **Emily Jin**, Jiajun Wu, Niloy Mitra. "FactoredScenes: Real-World Scene Generation via Library Learning of Room Structure and Object Pose Prediction." *Submitted at CVPR* (2025)
- [7] Wu, Sarah A., Erik Brockbank, Hannah Cha, Jan-Philipp Fränken, **Emily Jin**, Zhuoyi Huang, Weiyu Liu, Ruohan Zhang, Jiajun Wu, and Tobias Gerstenberg. "Whodunnit? Inferring What Happened from Multimodal Evidence." *In CogSci* (2024).
- [8] Huh, Minyoung, Brian Cheung, Tongzhou Wang, and Phillip Isola. "The Platonic Representation Hypothesis." arXiv preprint arXiv:2405.07987 (2024).
- [9] Sundaram, Shobhita, Stephanie Fu, Lukas Muttenthaler, Netanel Y. Tamir, Lucy Chai, Simon Kornblith, Trevor Darrell, and Phillip Isola. "When Does Perceptual Alignment Benefit Vision Representations?." *In NeurIPS* (2024).
- [10] Feng, Chun*, Joy Hsu*, Weiyu Liu, and Jiajun Wu. "Naturally Supervised 3D Visual Grounding with Language-Regularized Concept Learners." *In CVPR* (2024).
- [11] Hsu, Joy, Jiayuan Mao, Josh Tenenbaum, and Jiajun Wu. "What's Left? Concept Grounding with Logic-Enhanced Foundation Models." *In NeurIPS* (2024).
- [12] Liu, Weiyu, Geng Chen, Joy Hsu, Jiayuan Mao, and Jiajun Wu. "Learning Planning Abstractions from Language." *In ICLR* (2024).
- [13] Chen, Allison, Ilia Sucholutsky, Olga Russakovsky, and Thomas L. Griffiths. "Analyzing the roles of language and vision in learning from limited data." arXiv preprint arXiv:2403.19669 (2024).
- [14] Hu, Yushi, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. "Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models." *In NeurIPS* (2024).
- [15] Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. "Building Machines that Learn and Think Like People." *Behavioral and Brain Sciences* 40 (2017): e253.