Statement of Purpose

zcczhang.github.io **Zichen "Charles" Zhang** *charlesz@allenai.org*

I am interested in both **generalist** and **specialist** embodied agents. My ambition is to build a **unified, general-purpose, agent** capable of understanding and engaging in the multimodal world. Such an agent should be capable of mobile manipulation with 3D interactions, assisting humans in the real world from multi-sensory inputs, and lifelong autonomous and adaptive learning. To achieve these capabilities and further advance the field, I am eager to delve deeper into cutting-edge research by pursuing a Ph.D. degree. In pursuit of my long-term goals, I have **led** or **co-led** a number of projects [1, 2, 3] as a Predoctoral Young Investigator at the Allen Institute for AI (AI2), under the supervision of Luca Weihs and Jiasen Lu. Before joining AI2, I collaborated with Linxi "Jim" Fan and Yuke Zhu [4], and led summer research and research internship projects [5, 6, 7] during my undergraduate studies at Macalester College. These experiences laid a strong foundation for my research and kindled my fascination with it.

Autonomous RL (ARL) with Minimal Interventions. We expect intelligent agents to autonomously learn with external interventions *only* when necessary. Previous reset-free RL works are promising (e.g. [8, 9]), but require carefully designed environments and face challenges arising from limited state diversity. These challenges confined previous studies to visually simplistic environments that do not require substantial generalization during inference. In [3], I first identified that a main challenge for Autonomous RL (ARL) is "**near-irreversible states**". These are states that are not strictly irreversible but can pose considerable difficulty for an agent, such as when an object is pushed to the edge of its reach in early training stages from which recovery is unlikely. I developed **unsupervised distance-** and **dispersion-based measures** to identify when irreversible states have been reached during training. These measures allowed us to solve the well-known SawyerPeg task in ARL (related previous work achieved near-0% success).

Beyond the proposed methodology, in this work I developed a **pick-and-place mobile manipulation benchmark** within the high visual fidelity simulator AI2-THOR [10], designed to assess four levels of generalization abilities. Furthermore, by noting that an agent training without interruptions might, intuitively, be expected to explore more deeply than one trained with episodic RL, I introduced a **novel ARL framework** which expands the agent's goal space to include the majority of the agent's state space. Our methods also mark the first foray of **ARL into the embodied navigation space**: we achieved a higher success rate and used 99.97% fewer resets than an episodic object navigation baseline in RoboTHOR [11].

Pretrained Visual Representations Enable Task Decomposition. The achievements of vision (& language) backbones for decision-making have led to a focus on short-horizon tasks (e.g. [12, 13]), while recent approaches for long-horizon manipulation typically rely on LLMs or generative models, requiring additional domain-specific data for pretraining. My insights from ARL, and the discussions with the authors of [13] at NeurIPS 2022, led me to believe that distances in feature space from pretrained visual representations can effectively indicate state phase changes of RGB videos. This idea culminated in the Universal Visual Decomposer (UVD) algorithm [2], which identifies subgoals by **recursively detecting phase shifts** within a given video demonstration. Significantly, UVD operates entirely **off-the-shelf**, requiring **no extra data**, **training**, or **task knowledge**, and can be integrated with **any** visuomotor algorithms.

Since UVD decomposes subgoals from any unlabeled trajectory, I designed and verified how these subgoals can be used to enable **compositional generalizations** in multitask learning and the creation of **monotonic RL rewards**. By training a goal-conditioned behavior-cloning (GCBC) agent on partial sub-task sequences in FrankaKitchen, I showed that UVD subgoals enable generalization to unseen combinations, across **various visual backbones and policy architectures**. I also showed that using UVD-rewards for RL successfully enabled learning multi-stage, long-horizon, manipulation tasks **without reward engineering.** UVD was successfully applied in the real world where plug-and-play UVD substantially improved out-of-domain generalization. As the lead author, I was thrilled that UVD won the **Best Paper** award at the Learning Effective Abstractions for Planning (LEAP) workshop at CoRL'23, and UVD was selected for **oral presentations** at LEAP and Foundation Models for Decision Making (FMDM@NeurIPS'23) workshops.

Multimodal Prompting Formulation for Robot Manipulation. A generalist robot should have an intuitive and expressive interface for task specification. With collaborations, I contributed to developing and improving VIMA-Bench [4], the first benchmark supporting 17 representative manipulation tasks guided by

multimodal instructions like text, image, and video. In developing VIMA-Bench (built using PyBullet and extended from CLIPort [14]), I addressed several key limitations, e.g. insufficient consideration of distractors, and made more task variations for testing generalization. I also streamlined the task suite with more efficient, unified APIs. I was involved in the model implementation, initial experiments and ablations, and in developing detection modules for object-centric manipulation, all contributing to the creation of the novel **Vi**suo**M**otor **A**ttention (VIMA) model.

This was a valuable collaboration with experts from Stanford, NVIDIA, UT-Austin, and Caltech, and later, I also contributed to developing skill primitives in MineDojo [15] (precursor of Voyager [16]), and an unreleased project for few-shot imitation learning where I sped up the distributed training infrastructure and proposed the idea of using contrastive learning for analogy-making.

Modalities and Tasks Unification. We recently released Unified-IO 2 [1], the **first** model that enables understanding and generation across vision, language, audio, video, and action. Joining the project after its start, I initially focused on incorporating embodied benchmarks (e.g. VIMA, Habitat). I was able to significantly expand the scope of my contributions, leading to my role as a **co-leading** author, by rapidly learning TensorFlow and Jax, building upon T5X and SeqIO frameworks, and running TPU training. Training a multimodal autoregressive model from **scratch** to **generate text, image, and audio with sparse and dense inputs/outputs** was challenging, relying heavily on careful engineering. I identified and fixed many hidden issues in the data, model, and training processes. My contributions also included adding robot, audio, video, and multimodal instruction tuning datasets, along with implementing novel **augmentations for all modalities** in both pretraining and instructional finetuning phases, culminating in over 230 tasks. Notably, the action representation and embodied question-answering augmentations I introduced predated recent releases of RT-2 [17] and RoboVQA [18]. I conducted numerous experiments that led to **stabilization techniques** for training, particularly addressing instabilities from perceiver resamplers. During the project's early stages, these frequent adjustments based on extensive ablations, highlighted the surprisingly significant impact that even minor tweaks and augmentations could have. Moreover, I contributed to probing the model's **emergent abilities**. It was remarkable to observe that a model trained to master numerous tasks could also excel at **novel instruction following**, particularly in vision perception, image generation, and editing. This research provided me with invaluable experience with training 7B models from scratch using more than 600TB scale data that can greatly benefit other and future works, even with smaller scales. For instance, I brought rotary encoding and normalization techniques to the UVD project, leading to better stability during visuomotor training. Diverse modalities and tasks also introduced me to previously unfamiliar fields, broadening my perspectives and fueling my ambition to contribute towards the future of AI.

Future Directions. I plan to integrate the recent large-scale, mobile manipulation benchmark Chores [19] at AI2 into the Open X-Embodiment dataset [20]. Extending Unified-IO 2 for more specialized embodied agents, this effort aims to assess how a diverse generalist model, potentially acting as an implicit or explicit world model, influences large-scale decision-making and emergent capabilities across various tasks and embodiment in sim and real. I am also part of a, company-wide, collaboration to create an open-source GPT4-V level generalist at AI2. Unlike the Unified-IO 2 model which was a multi-modality generalist, this new project aims to maximize performance with a more narrow scope of vision and language capabilities.

Generalists are appealing, but there remain a myriad of exciting challenges and directions that require specialization. I wish to build specialized generalists that are **embodied** in real robots and use multimodal sensory inputs to interact with humans. Additionally, the success of Diffusion Policies in robotics and the effectiveness of diffusion/consistency models in dense generations (depth, segmentation, optical flow, etc.), as well as in 3D and video domains, leads me to think about how features from diffusion models could further bridge the gap between perception and action. Since both the world and embodied AI are inherently 4D and multimodal I believe there's potential for groundbreaking insights in this intersection. My rich research and engineering experience with large-scale multimodal model and embodied AI will be crucial in exploring these possibilities.

References

[1] Jiasen Lu*, Christopher Clark*, Sangho Lee*, **Zichen Zhang***, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. "Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action". *Under Review* (2023).

[2] **Zichen Zhang**, Yunshuang Li, Osbert Bastani, Abhishek Gupta, Dinesh Jayaraman, Yecheng Jason Ma, and Luca Weihs. "Universal Visual Decomposer: Long-Horizon Manipulation Made Easy". *Learning Effective Abstractions for Planning (LEAP) workshop (**Best Paper Award**), CoRL 2023. Foundation Models for Decision Making (FMDM) workshop (**Oral**, **6/112**), NeurIPS* (2023). arXiv: 2310.08581.

[3] **Zichen Zhang** and Luca Weihs. "When Learning Is Out of Reach, Reset: Generalization in Autonomous Visuomotor Reinforcement Learning". *Out-of-Distribution Generalization in Robotics workshop (**Lightning Talk**), CoRL* (2023). arXiv: 2303.17600.

[4] Yunfan Jiang, Agrim Gupta*, **Zichen Zhang***, Guanzhi Wang*, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. "VIMA: General Robot Manipulation with Multimodal Prompts". *Foundation Models for Decision Making (FMDM) workshop (**Oral**), NeurIPS 2022. Proceedings of Fortieth International Conference on Machine Learning (ICML)*. 2023.

[5] **Zichen Zhang**, Yutong Wu, and Lisa Naples. "Characterization of Rectifiable Measures Carried by Lipschitz Curves". *Joint Mathematics Meeting (JMM)* (2022).

[6] **Zhang Zichen**, Lang Wang, and Shuhao Wang. "Automated Scoring System of HER2 in Pathological Images under the Microscope". *18th European Congress on Digital Pathology (ECDP)* (2022).

[7] **Zhang Zichen**, Fan Zhang, Elisabeth Landgren, Aaron Gould, and Esra Kadioglu Urtis. *Area Coverage with Unmanned Aerial Vehicles Using Reinforcement Learning*. Technical Report. Macalester College, 2020.

[8] Eysenbach, *et al.* "Leave no trace: Learning to reset for safe and autonomous reinforcement learning". *ICLR*. 2018.

[9] Sharma, *et al.* "Autonomous Reinforcement Learning: Formalism and Benchmarking". *ICLR*. 2022.

[10] Kolve, *et al.* "AI2-THOR: An Interactive 3D Environment for Visual AI". *arXiv* (2017).

[11] Deitke, *et al.* "Robothor: An Open Simulation-to-Real Embodied AI Platform". *CVPR*. 2020.

[12] Nair, *et al.* "R3M: A Universal Visual Representation for Robot Manipulation". 2022.

[13] Ma, *et al.* "VIP: Towards Universal Visual Reward and Representation via Value-Implicit Pre-training". 2022.

[14] Shridhar, *et al.* "CLIPort: What and Where Pathways for Robotic Manipulation". *CoRL*. 2022.

[15] Fan, *et al.* "MineDojo: Building Open-ended Embodied Agents with Internet-scale Knowledge". 2022.

[16] Wang, *et al.* "Voyager: An Open-ended Embodied Agent with Large Language Models". *arXiv* (2023).

[17] Brohan, *et al.* "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control". *CoRL*. 2023.

[18] Sermanet, *et al.* "RoboVQA: Multimodal Long-Horizon Reasoning for Robotics". *arXiv* (2023).

[19] Ehsani, *et al.* "Imitating Shortest Paths in Simulation Enables Effective Navigation and Manipulation in the Real World". *arXiv* (2023).

[20] Padalkar, *et al.* "Open X-Embodiment: Robotic Learning Datasets and RT-X models". *arXiv* (2023).