Xuweiyi Chen

xuweic@email.virginia.edu | (206)532-9635 | xuweiyichen.github.io

EDUCATIONAL BACKGROUND

UNIVERSITY OF VIRGINIA **Charlottesville, VA**
Ph.D. in Computer Science and Engineering *Aug. 2024 – May 2029 (Expected)*
Overall GPA: 4.0/4.0
Concentration: **3D Computer Vision and Robot Learning**

UNIVERSITY OF MICHIGAN **Ann Arbor, MI**
M.S. in Computer Science and Engineering *Aug. 2022 – May 2024*

UNIVERSITY OF WASHINGTON **Seattle, WA**
B.S. in Applied and Computational Mathematical Sciences, CUM LAUDE *Sep. 2018 – June 2022*
Honors: $6000 CoMotion Mary Gates Innovation Scholarship
$3000 Usha and S. Rao Varanassi SAFS Scholarship

SELECTED INTERNSHIPS

Lambda **San Francisco**
Machine Learning Research Intern. *Jan. 2025 – May. 2025*

- Experience Large-scale Pretrained Multi-modal Models using 24 B200 GPUs
- Designed novel architectures for multi-modal unifying 2D, 3D and 4D all using one single latent vector.
- Led the integration of computer vision with other modalities by developing unified multimodal representations that enable joint reasoning across language, vision, and audio.

SELECTED FIRST-AUTHOR PUBLICATIONS

Semantic-Free Procedural 3D Shapes Are Surprisingly Good Teachers **3DV 2026**
UVA CV LAB supervised Prof. Zezhou Cheng *Nov. 2024*

- Procedurally generated shapes offer a scalable, copyright-free, and geometrically diverse alternative to labor-intensive human-designed 3D datasets like ShapeNet.
- We use procedurally generated 3D shapes to achieve strong results in object classification, part segmentation, and few-shot learning.
- Point-MAE-Zero can perform masked point cloud completion without fine-tuning.

Probing the Mid-level Vision Capabilities of Self-Supervised Learning **CVPR 2025**
UVA CV LAB supervised Prof. Zezhou Cheng *Nov. 2024*

- Developed a benchmark suite to systematically evaluate mid-level vision capabilities in SSL models across 8 tasks.
- Conducted a large-scale study assessing 22 SSL models, revealing weak correlations between mid-level and high-level vision performance.
- Identified key factors influencing mid-level vision performance, including pretraining objectives and network architectures, providing insights for future SSL research.

3D-GRAND: A Million-Scale Dataset for 3D-LLMs with Better Grounding and Less Hallucination **CVPR 2025**
SLED lab in the University of Michigan supervised Prof. Joyce Chai & Prof. David Fouhey *Aug. 2024*

- Introduced 3D-GRAND, a large-scale dataset with 40,087 household scenes and 6.2 million densely grounded scene-language instructions to improve 3D-Language models (3D-LLMs).
- Proposed 3D-POPE, a benchmark to evaluate hallucinations in 3D-LLMs, enabling fair comparisons across models.
- Demonstrated that instruction tuning with 3D-GRAND significantly enhances grounding capabilities, emphasizing the importance of large-scale 3D-text datasets for advancing embodied AI research.

Multi-Object Hallucination in Vision-Language Models **NeurIPS 2024**
SLED lab in the University of Michigan supervised Prof. Joyce Chai & Prof. David Fouhey *July 2024*

- Investigated multi-object hallucination in Large Vision Language Models (LVLMs) using Recognition-based Object Probing Evaluation (ROPE), focusing on the distribution of object classes within a single image and visual referring prompts.
- Found that LVLMs exhibit more hallucinations when tasked with recognizing multiple objects compared to a single object, influenced by object class distribution and model behaviors.
- Identified key factors such as salience, frequency, and model intrinsic behaviors that contribute to hallucination, aiming to improve LVLMs' recognition and reasoning capabilities in complex visual scenes.

LLM-Grounder: Open-Vocabulary 3D Visual Grounding with Large Language Model as an Agent. **ICRA 2024**
SLED lab in the University of Michigan supervised Prof. Joyce Chai *Aug. 2023*

- Present the first method capable of localizing novel objects in 3D scenes using Neural Radiance Field (NeRF) and Large Language Models (LLMs) through iterative, natural language-based interactions.
- Enables a more human-like interaction with 3D objects in a learned 3D scene representation.
- Evaluated and shown that dynamic grounding outperforms static grounding in terms of accuracy, 3DIoU, and human ratings.