Yunong Liu

Palo Alto, CA | yunong.liu20@gmail.com | yunongliu.com | github.com/yunongLiu1 | Google Scholar | x.com/yunongliu1

Summary

Research Engineer at Luma AI working across image/video generative modeling and model improvement: diffusion / multimodal modeling, training and evaluation workflows, distillation, reward/verifier design, failure analysis, and model selection. Built Ray3 video-generation modeling / evaluation workflows and GRPO / DPO-style post-training for Ray3 and Uni-1; designed OCR, aesthetic, motion, physics, and rubric / VLM rewards; and lead Layering, a structured multi-layer generation project for editable visual scenes. Previously first author on Stanford 4D grounding work aligning real internet assembly videos with 3D models, instructions, and temporal task structure.

Education

Stanford University Sep 2023 – Apr 2025

M.S. in Computer Science, GPA 3.97 / 4.00, Distinction in Research Stanford, CA Advised by Prof. Jiajun Wu. Focus: vision, multimodal learning, generative models, 4D grounding, and model evaluation.

Selected coursework: CS381 Sensorimotor Learning for Embodied Agents; CS326 Topics in Advanced Robotic Manipulation; CS348I Computer Graphics in the Era of AI; CS148 Introduction to Computer Graphics and Imaging.

The University of Edinburgh Sep 2019 – Jun 2023

BEng in Electronics and Computer Science, Joint Honours; ranked 2nd in cohort Edinburgh, UK

The University of Texas at Austin Jan 2022 – Jun 2022

Visiting student, Electrical and Computer Engineering; GPA 3.82 / 4.00 Austin, TX

Core Expertise

Research: large image and video generative models; video generation modeling; multimodal generation; OCR and aesthetic reward modeling; physical / temporal consistency evaluation; 4D grounding; VLM-as-judge evaluation.

Modeling: diffusion transformers; Transfusion / MMDiT-style multimodal modeling; structured multi-layer / multi-image generation; native-RGBA latent modeling; RL, GRPO-style, and DPO-style post-training; reward hacking and over-optimization analysis.

Engineering: Python; PyTorch; model training and evaluation workflows; sample-generation harnesses; systematic ablations and checkpoint selection; reward / verifier pipelines; data filtering for model improvement.

Experience

Luma AI Apr 2025 – Present

Research Engineer – image/video generation, post-training, model improvement Palo Alto, CA

- **Video generation modeling.** Implemented an LTX-Video-style causal video autoencoder / tokenizer in a production training framework, with causal 3D convs, pixel-shuffle / pixel-norm components, dual Conv3D paths, and 121-frame 512px reconstruction experiments.
- Built Ray3 video-modeling and post-training workflows spanning sample generation, reward training, VLM-as-judge grading, held-out benchmarking, failure analysis, and checkpoint selection.
- Designed video data and evaluation experiments around temporal labels, cut / fade boundaries, style, prompt / caption distributions, and FPS-normalized frame-boundary handling for cleaner model-improvement loops.
- **RL / post-training.** Integrated GRPO-style training into the main diffusion-model framework and ran diffusion RL, DPO-style, Diffusion-NFT, and SFT / RL alternation experiments for image and video generators.
- Designed reward / verifier signals across prompt following, OCR, aesthetic quality, motion, composition, and physical plausibility, using learned preference rewards, OCR rewards, and rubric / VLM judges; studied reward hacking, oversaturation, large action-space failures, and held-out human-preference calibration.
- **Uni-1 – unified multimodal foundation model.** Contributed to Uni-1 post-training, RL / GRPO experiments, reward modeling, evaluation, and model selection for unified image understanding and generation.

- **Layering – structured multi-layer image generation.** Project lead for a model that decomposes and generates visual scenes as ordered editable components – raster objects, text, vectors / SVG, alpha, and z-order – rather than a flat image.
- Trained a native-RGBA VAE and spec-conditioned layered diffusion model for editable visual elements conditioned on layout, style, text, alpha, and z-order; designed decomposition, object extraction, add-layer generation, and layer-conditioned reconstruction tasks for structured scene factorization.
- Built reconstruction, VLM-as-judge, and semantic evaluation checks for modeling failures pixel metrics miss, including incomplete objects, cross-layer duplicates, prompt adherence, color fidelity, text grounding, alpha quality, and layer editability.

Stanford Vision and Learning Lab Jun 2023 – Apr 2025
Research Assistant with Jiajun Wu, Juan Carlos Niebles, Manling Li, Weiyu Liu, Cristobal Eyzaguirre Stanford, CA

- Sole student lead and first author on **IKEA Manuals at Work**, a year-long NeurIPS 2024 Datasets and Benchmarks project aligning real internet assembly videos with 3D models, instruction manuals, and temporal grounding annotations to recover object / action / task structure.
- Built cross-frame optimization combining PnP-RANSAC with temporal-consistency constraints across **34k+ frames** from **98 videos**; coordinated **30 annotators** with iterative validation and quality control.
- Contributed to zero-shot optical-flow extraction from video diffusion models by probing counterfactual changes in video-diffusion logits, yielding state-of-the-art TAP-Vid performance without labels or fine-tuning.
- Awarded **Distinction in Research**; completed four quarters of funded RA work while maintaining **3.97 / 4.00** GPA.

University of Edinburgh Jul 2022 – Jan 2023
Research Assistant – discourse relation analysis with Prof. Bonnie Webber Edinburgh, UK

- Worked on NLP research for discourse relation analysis, building experience in language annotation, model evaluation, and empirical error analysis.

Selected Publications

CaptionQA: Is Your Caption as Useful as the Image Itself? CVPR 2026
Shijia Yang*, **Yunong Liu***, Bohan Zhai*, Ximeng Sun, Zicheng Liu, Emad Barsoum, Manling Li, Chenfeng Xu
Equal contribution. Utility-based multimodal benchmark with 33,027 multiple-choice questions across natural, document, e-commerce, and embodied-AI domains.
arXiv | Code

Taming Generative Video Models for Zero-Shot Optical Flow Extraction NeurIPS 2025
Seungwoo Kim*, Khai Loong Aw*, Klemen Kotar*, Cristobal Eyzaguirre, Wanhee Lee, **Yunong Liu**, Jared Watrous, Stefan Stojanov, Juan Carlos Niebles, Jiajun Wu, Daniel L. K. Yamins
Counterfactual probe over video-diffusion logits for optical flow with no labels and no fine-tuning; generalizes to in-the-wild videos and outperforms specialized baselines.
arXiv | Code | Project

IKEA Manuals at Work: 4D Grounding of Assembly Instructions on Internet Videos NeurIPS 2024 Datasets and Benchmarks
Yunong Liu, Cristobal Eyzaguirre, Manling Li, Shubh Khanna, Juan Carlos Niebles, Vineeth Ravi, Saumitra Mishra, Weiyu Liu*, Jiajun Wu*
First-author work and sole student lead. Dataset and evaluation framework aligning real assembly videos with 3D models and instruction manuals across 34k+ frames and 98 videos.
arXiv | Project

COVID-19 Misinformation Detection: Machine-Learned Solutions to the Infodemic JMIR Infodemiology 2022
Yunong Liu*, Nikhil Kolluri*, Dhiraj Murthy
Co-first author. Hybrid framework combining classical and pretrained language models with crowdsourced annotations for misinformation detection.
Paper