

RESEARCH ENGINEER · LUMA AI

Yunong Liu.

yunongliu.com
yunong.liu20@gmail.com
X / @yunongliu1
Palo Alto, CA · Updated 2026

Research Engineer at Luma AI working on *structured visual generation, post-training, and reward modeling.*

I lead *Layering*, a system that turns flat visual generations into editable raster, text, and vector components that humans and agents can revise across multiple turns. My work spans native-RGBA generative modeling, structured layer plans, verifier-driven editing loops, data pipelines, and evaluation. I also built the Ray3 video RL workflow end to end and contributed to Uni-1 post-training / RL experiments.

Stanford MSCS, *Distinction in Research*, advised by Jiajun Wu. First-author NeurIPS 2024, co-first CVPR 2026, contributor NeurIPS 2025, Edinburgh BEng, ranked 2nd of cohort.

🎓 I - EDUCATION

Education.

2019 - 2025

M.S. Computer Science	Stanford University M.S. Computer Science GPA 3.97 / 4.00 · Distinction in Research Advised by Prof. Jiajun Wu. Led 1 first-author (IKEA Video Manuals, NeurIPS '24) and 2 collaborative projects spanning multimodal evaluation and emergent structure in video-diffusion models.	Sep 2023 - Apr 2025 · Stanford, CA
BEng Electronics & CS	The University of Edinburgh BEng Electronics & Computer Science · Joint Honors · Ranked 2nd of cohort	Sep 2019 - Jun 2023 · Edinburgh, UK
Visiting ECE	The University of Texas at Austin Electrical & Computer Engineering · 18 credits / semester	Jan - Jun 2022 · Austin, TX

🔬 II - EXPERIENCE

Research *experience.*

2023 - present

2025 April - present
Palo Alto, CA

Luma AI · Research Engineer (Project Lead)

Generative models · multimodal · agentic systems

- **Layering — project lead.** Built *editable layered image generation*: convert flat 2K image/design outputs into ordered raster / text / vector layer stacks that humans and agents can revise over multiple turns. Owned model, data, eval, and agent loop end to end.
 - **Model.** Trained a **native-RGBA VAE** and **spec-conditioned layered diffusion model** that generates individual editable assets from explicit layer specs: layout, style, text, alpha, and z-order. Reframes image generation from flat pixels into structured composition, reducing cross-layer duplication and enabling targeted edits.
 - **CFG distillation.** Distilled a classifier-free-guided layered diffusion teacher into a **single-pass conditional student**, reducing inference compute while preserving layer quality.
 - **Eval.** Built **reconstruction**, **VLM-as-judge**, and **semantic** checks for failures pixel metrics miss: half-object layers, cross-layer duplicates, prompt adherence, color fidelity, text grounding, and layer editability.
 - **Agent loop.** Built a **VLM orchestrator** that parses free-form edits into typed layer operations and runs evals as **verifier-in-the-loop** checks, turning image generation into multi-turn *plan-execute-verify* rather than one-shot inference.
 - **Data.** Built **rendering pipelines** for PSD / Canva / Figma sources plus synthetic data. Designed **subset-layer augmentation** to turn each composite into many decomposition, extraction, and add-layer training pairs.



Layering - pipeline overview

- **Uni-1 — unified multimodal foundation model.** Worked on **post-training and evaluation** for a unified multimodal generator. Designed **diffusion RL / data ablation studies** and **caption-variance experiments** to measure how reward design, prompt distribution, and data mix shift generation quality, controllability, and robustness.
- **Ray3 — video generation RL.** Led the **RL experimentation stack** for Ray3 video generation, from data construction to reward training and held-out evaluation. Built **VLM-as-judge graders** and ran model-improvement loops across prompt following, motion, composition, and visual quality.
- **Reward modeling.** Built **calibrated reward / eval systems** for image and video generators, combining **learned preference models** with **rubric-based VLM judges** and held-out human-preference checks to reduce reward hacking while improving generation quality.

Stanford Vision & Learning Lab · Research Assistant

with Jiajun Wu, Juan Carlos Niebles, ManLing Li, Weiyu Liu, Cristobal Eyzaguirre

- **IKEA Video Manuals — NeurIPS 2024 (first author)**. Sole student lead on a year-long project: first dataset that aligns real-world assembly videos with 3D models in space and time. Built the cross-frame optimization combining PnP-RANSAC with temporal-consistency constraints; coordinated 30 annotators across 34k+ frames over 98 videos with iterative cross-validation; defined the evaluation protocol now used by follow-up 4D-grounding work.
- **Zero-shot optical flow from video diffusion — NeurIPS 2025 (contributor)**. Co-developed a counterfactual probe over video-diffusion logits that yields state-of-the-art TAP-Vid optical flow with no labels and no fine-tuning; generalizes to in-the-wild videos and outperforms specialized baselines.
- **Distinction in Research** · 4 quarters of funded RAship.

Selected *publications*.

2026

CVPR

CaptionQA — Is your caption as useful as the image itself?

S. Yang*, **Y. Liu***, B. Zhai*, X. Sun, Z. Liu, E. Barsoum, M. Li, C. Xu

Utility-based caption benchmark — 33k MCQs across Natural / Document / E-commerce / Embodied; surfaces image-caption utility gaps standard QA-on-image benchmarks miss.

[arXiv 2511.21025](#) [github](#) / [bronyayang](#) → [paper](#)

2025

NeurIPS

Taming generative video models for *zero-shot optical flow* extraction

S. Kim*, K. L. Aw*, K. Kotar*, C. Eyzaguirre, W. Lee, **Y. Liu**, J. Watrous, S. Stojanov, J. C. Niebles, J. Wu, D. L. K. Yamins

Counterfactual probe over video-diffusion logits — SOTA TAP-Vid optical flow with no labels and no fine-tuning.

[arXiv 2507.09082](#) [code](#) / [kl_tracing](#) → [project page](#)

2024

NeurIPS

Datasets &

Benchmarks Track

IKEA Manuals at Work — 4D grounding of assembly instructions on internet videos

Y. Liu, C. Eyzaguirre, M. Li, S. Khanna, J. C. Niebles, V. Ravi, S. Mishra, W. Liu*, J. Wu*

First dataset aligning real-world assembly videos with 3D models — 34k+ frames, 98 videos. Sole student lead, year-long project; coordinated 30-annotator cross-validation pipeline.

[arXiv 2411.11409](#) [page](#) / [ikea-video](#) → [project page](#)

What I'm *building recently*.

Interactive visual generation.

LUMA AI · LAYERING · PROJECT LEAD

Most image and video generators are still *single-turn* systems: a prompt goes in, a flat raster comes out, and meaningful edits usually require re-rolling the whole result. Real creative work is different. It is iterative, spatial, and collaborative: people point, revise, and refine; agents inspect the canvas, propose changes, and verify whether the result matches intent.

Layering is my current bet on making visual generation *interactive by design*. Instead of treating the output as an opaque pixel grid, the model produces an *editable stack of raster, text, and vector components*. These components can be inspected, rearranged, regenerated, and checked over multiple turns, turning generation from a one-shot sample into a *plan-execute-verify* loop.

The broader direction is *generation beyond pixels*: models should emit structured components where structure matters, pixels where visual richness matters, and interfaces that let humans and agents work with both.

RESEARCH INTERESTS

I'm interested in generative systems whose outputs are inspectable, editable, and verifiable, especially where visual generation connects pixels to structured or code-backed representations.

STRUCTURED VISUAL OUTPUT

Generating **layers, regions, components, and layouts** that can be read and revised by humans and agents, instead of opaque rasters that can only be re-sampled.

REWARD & POST-TRAINING

RLHF, reward modeling, and calibration for image and video generators, with an emphasis on reward signals that improve controllability without collapsing into reward hacking.

LONG-HORIZON MULTIMODAL EVAL

Benchmarks and **verifier loops** that measure whether generated outputs remain useful across multi-turn interactions, not just whether a single response looks good.