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* — turning flat pixels into editable raster, text, and vector components that humans and agents revise across *multiple turns*, combining generation with **code, actions, and tools**. Also built the Ray3 RL workflow and contributed to Uni-1 RL.

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.** Turn flat 2K image generations into *ordered, editable RGBA layer stacks* that humans and agents can revise across multiple turns. End-to-end ownership across model, eval, agent loop, and data.

Layering — pipeline overview

  - **Model.** Trained a **native-RGBA VAE** with a distillation objective against the base RGB VAE to preserve the image prior. Designed the **diffusion conditioning** for layered generation — each layer is synthesized conditioned on an explicit per-layer specification rather than all layers at once, which avoids duplicate or overlapping content across layers. **Crop-based layer tokenization** fits 20+ layers in 2K context (vs ~4); training distributed across **8–32 nodes** per run.
  - **Eval.** Three axes: **reconstruction** (PSNR / SSIM), **VLM-as-judge global metrics** (prompt adherence, color fidelity, text grounding), and **semantic-correctness** checks for failures pixel metrics miss — half-object layers, cross-layer duplicates.
  - **Agent loop. VLM orchestrator** parses free-form edits into typed layer ops and runs the eval suite as a **verifier-in-the-loop** check. Image generation as multi-turn *plan-execute-verify* rather than one-shot inference.
  - **Data.** Built **rendering pipelines** for **PSD / Canva / Figma** sources and synthetic data; **subset-layer augmentation** that turns each composite into many decompose / extract / add training pairs.

- **Uni-1 — unified multimodal foundation model.** Contributed **RL** and data experiments toward unified multimodal understanding & generation.

- **Ray3 — video generation.** Built the Ray3 **RL workflow** end-to-end: data pipelines, reward training (incl. **VLM-as-judge graders**), and held-out evaluation, with training runs distributed across **8–32 nodes**. Ran experiments across video quality, prompt following, motion, and composition.

- **Reward modeling & calibration.** *Preference-aligned* reward + calibration experiments across image and video generators; trained reward models combined with rubric-prompted **VLM judges**, calibrated against held-out human preference and designed to resist reward over-optimization, covering prompt following, composition, aesthetics, and motion.

2023 Jun — 2025 April
Stanford, CA

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.

III — RESEARCH Selected *publications.*

2024 — 2026

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

IV — CURRENTLY What I'm *building recently.*

LUMA · LAYERING
2026 —

Interactive visual generation.

LUMA AI · LAYERING · PROJECT LEAD

Most of today's image and video generators are effectively *single-turn*: a prompt goes in, a flat raster comes out, and the only way to change anything is to re-roll the whole thing. But real creative work is *multi-turn* — a conversation that unfolds over many revisions. People want to *point, edit, and revise* mid-stream; agents want to read the canvas, propose a change, and verify the result. Neither is possible when the output is an opaque pixel grid.

My current interest is making generative visual models *interactive by design* — outputs that survive being edited, redirected, and refined across many turns, rather than collapsing into a re-roll. **Layering** is my first big bet in this direction: a structured output that humans and agents can both read, edit, and verify against intent — turning generation from a one-shot guess into a *long-horizon* loop. The deeper bet is that generation should reach beyond pixels — emitting *code-backed components* where structure matters, and pixels where visual richness matters.

RESEARCH INTERESTS

I'm interested in generative systems whose outputs are inspectable, editable, and verifiable. Especially where pixel outputs connect to structured or code-backed representations.

REWARD	**RLHF** and reward modeling for image and video generators; calibration against held-out human preference; reward models that resist over-optimization.
STRUCTURED OUTPUT	Generation as **structured output** (layers, regions, components) that agents can read and revise, not opaque rasters that can only be re-rolled.
LONG-HORIZON EVAL	Utility-based **multimodal benchmarks** (CaptionQA, IKEA Manuals) that measure whether outputs remain useful across **multi-turn** interactions — not just a single response.