Jiahang He

Berkeley, CA
(669) 212-0469 | jhe.primary@gmail.com

EDUCATION:

Regents & Chancellor's Scholar at **UC Berkeley**
B.S. Data Science | Expected Graduation: Aug 2027
GPA: 4.0

RESEARCH PUBLICATIONS:

First Author, *"EGO-FLIGHT: Egocentric Grounding of Order for Frame-Level Inference in General Human Timelines."* **ICLR 2026 Workshop on World Models: Understanding, Modelling and Scaling**

- Built benchmark of 1,056 egocentric video clips evaluating temporal reasoning in VLMs through frame-ordering tasks across 4 controlled variables.
- Demonstrated that LM-only LoRA fine-tuning improves temporal ordering accuracy from τ=0.03 to τ=0.17 on 4-frame clips with strong initial positioning anchor effects.

First Author, *"Modeling and Predicting Multi-Turn Answer Instability in Large Language Models."*
NeurIPS 2025 Workshop on Multi-Turn Interactions

- Revealed up to 34.8% stationary accuracy degradation in Claude 3.5 Haiku under multi-turn prompting across 4 datasets, modeled via Markov chains predicting accuracy within 4% deviation.
- Probed hidden states of Gemma 3 4B to find answer changes detectable by layer 3, with linear probe accuracy rising from 0.50 to 0.89 within the first 3 layers.

First Author, *"Mechanistic Analysis and Inference-Time Control of Modality Conflict in VLMs."*
ICML 2026 Mechanistic Interpretability Workshop

- Identified the bridge module as the sole causal entry point for visual information in VLMs via contribution patching, and localized modality commitment to specific layers: LLaVA commits mid-network at layers 17–18 while Qwen flips late at layers 22–23.
- Developed a stateless MLP offset vector achieving bidirectional inference-time modality control without retraining, pushing LLaVA's vision-following rate from 9.2% to 38.8% and recovering Qwen's from 80.7% to 94.9% via targeted MLP replacement.

First Author, *"Subliminal Transfer of Positional Biases in Language Models."*
ICML 2026 Trustworthy AI for Good Workshop

- Demonstrated subliminal letter-bias transfer in 7 of 8 instruction-tuned LMs across 6 organizations (target-letter shifts up to ~10 pp), extending the subliminal learning framework from semantic to structural biases for the first time.
- Discovered a representation-behavior dissociation in Phi-3-medium where a clean, causally steerable bias direction localizes to a residual-stream direction at layer 17 yet produces zero behavioral letter selection, showing behavioral auditing alone cannot certify a distilled model is free of inherited biases.

First Author, *"An Evaluation of Vision-Language Models on Graph Reconstruction."*
IEEE VIS 2025 Workshop on GenAI

- Built a reverse-reasoning benchmark requiring VLMs to reconstruct full datasets from chart images across 8 chart types.
- Found that chart-generation capability does not predict reconstruction accuracy: GPT-4 generated all 8 chart types but had the highest reconstruction error of all models tested.

Co-author, *"WristCompass: Kinematic Coupling as a Learnable Visual Concept for Ego-Camera Orientation."* **CVPR 2026 Workshop on Visual Concepts**

- Designed and validated the 4D inter-wrist feature pipeline, demonstrating it outperforms 126D full hand keypoints across all 5 folds of subject-level cross-validation.
- Demonstrated zero-shot transfer from tabletop manipulation to Epic Kitchens cooking video at 14.3° median geodesic error, matching a 1B-parameter scene model at 1000× lower parameter count.

Co-author, *"Most of This Video is Boring."* **CVPR 2026 Workshop on Visual Concepts**

- Designed an event-driven video encoder that tokenizes only state-change transitions, achieving 200× compression over naive VLM tokenization with ~1.1M trainable parameters.

- Outperformed LongVU and uniform subsampling on GUI-World QA with an adaptive LAB-space transition detector at F1=0.939 and 90× encoding speedup over naive per-frame encoding.

Co-author, *"Before You Act: Benchmarking Interpretability in Vision-Language Models."* **COLM 2026 Actionable Interpretability Workshop**

- Built MMIB, a benchmark extending mechanistic interpretability evaluation (CPR/CMD, IIA) to VLMs, with a procedural CLEVR counterfactual dataset.
- Benchmarked CMA, V-SEAM, DAS, DBM, and PCA across BLIP, LLaVA, and Qwen2.5-VL; found LLaVA encodes visual input but lets text priors override it under conflict.

Co-author, *"Line Bring-Up Is a Coordination Problem: A System for Standing Up Multi-Cell Learned-Policy Production Lines."* **RSS 2026 Lab-to-Production Workshop**

- Built a sequential line bring-up system with a "propagation graph" quantifying how upstream drift induces downstream failures, surfacing inter-cell coupling invisible to per-cell tests.
- Raised line-level yield from 0.65 to 0.87 on a three-station OpenVLA-LIBERO line; showed upstream steering matches downstream fine-tuning at ~1/10 the cost.

RESEARCH EXPERIENCE:

Sky Computing Lab — UC Berkeley
Undergraduate Researcher | *May 2026 – Present*
Mentor: David Chan (Postdoctoral Researcher)

- Understanding how DINO's augmentations determine what its encoder learns and using discovered shortcut mechanisms to design better augmentation policies.

WORK EXPERIENCE:

Relling Systems (YC S25) | San Francisco, CA
Research Engineer (15 hours/week) | Nov 2025 – Aug 2026

- Scaling Laws Research
  - Developed scaling laws for robotics datasets by training hundreds of vision-language action models to identify optimal training data compositions.
  - Built an automated pipeline for annotation and task segmentation across thousands of hours of video data.
- Temporal Probing Research
  - Discovered and mechanistically traced temporal position bias in vision-language models, showing VLMs rely on input position as a proxy for chronological order.
  - Designed and ran a probing framework that revealed a stark dissociation between near-perfect input position encoding and near-chance temporal encoding.

Frizzle AI (YC S25) | San Francisco, CA
AI Intern (10 hours/week) | Dec 2025 – Aug 2026

- Built AI-powered math worksheet grading system
  - Combined vision-language models with OCR techniques to achieve pixel-precise error localization on handwritten math work.
  - Designed anchor-based matching pipeline that cross-validates OCR and LLM outputs to navigate printed vs. handwritten content for sub-character positioning.
  - Engineered adaptive whitespace detection system that positions LLM-generated feedback annotations at contextually appropriate locations.