Zoe W. He

https://zoewhe.github.io/ linkedin.com/in/zoewhe zoe.wy.he@gmail.com wah016@ucsd.edu

Education

University of California, San Diego Sept. 2020 – Present
Ph.D. in Cognitive Science
Advisor: Meenakshi Khosla

University of California, San Diego Sept. 2016 – June 2020
B.S. in Cognitive Science spec. Machine Learning & Neural Computation Summa Cum Laude
B.S. in Mathematics, Applied GPA: 3.95

Selected Publications & Manuscripts

Probing Fine-Grained Semantic Alignment Across Vision and Language with Minimal Pairs
***Zoe Wanying He**, Yanyu Lei, & Meenakshi Khosla*
Under review at ICLR 2027.

Barycentric Alignment for Instance-Level Comparison of Neural Representations
Shreya Saha, **Zoe Wanying He***, & Meenakshi Khosla*
Under review at ICLR 2027. [arXiv]
*Equal contribution.

Seeing Through Words, Speaking Through Pixels: Deep Representational Alignment Between Vision and Language Models
***Zoe Wanying He**, Sean Trott, & Meenakshi Khosla*
Oral Presentation.
2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025). [Paper]

Stability Implies Redundancy: Delta Attention Selective Halting for Efficient Long-Context Prefilling
*Yujie Chen, Tailai Chen, Yifeng Gao, **Zoe Wanying He**, Yijue Xu, Shaobo Wang, & Linfeng Zhang*
64th Annual Meeting of the Association for Computational Linguistics (ACL 2026). [Paper]

Pupil Size Reflects the Relevance of Reward Prediction Error and Estimation Uncertainty in Upcoming Choice
***Zoe W. He**, Maëva L'Hôtellier, Dalin Guo, Alexander Paunov, Florent Meyniel, & Angela J. Yu*
Oral Presentation.
46th Annual Meeting of the Cognitive Science Society (CogSci 2024). [Paper]

Multiple and Subject-Specific Roles of Uncertainty in Reward-Guided Decision-Making
*Alexander Paunov, Maëva L'Hôtellier, Florent Meyniel, Dalin Guo, **Zoe He**, & Angela J. Yu*
eLife Reviewed Preprint (2024). [Paper]

Technical & Professional Skills

Programming: Python, MATLAB, R, Java, C++, JavaScript, HTML/CSS, LaTeX
Machine Learning & AI: PyTorch, TensorFlow, scikit-learn, model training/fine-tuning, representation learning, representational alignment metrics, LLM/VLM evaluation, linear probing, ablations
Computational Modeling & Experimentation: Bayesian modeling, reinforcement-learning, eye-tracking, EEG
Collaboration & Communication: Cross-disciplinary collaboration with cognitive scientists, neuroscientists, and ML researchers; mentoring and co-supervising undergraduate and master's students (theses and capstone projects); teaching as TA and data science tutor (Python, Java, data structures)
Languages: English, Chinese (bilingual; full professional proficiency in both)

Research Projects

Benchmarking Fine-Grained Semantic Understanding Across Vision and Language 2026

- Developed a minimal-pair benchmark to test whether different models preserve fine-grained semantic distinctions in image-text matching across cognitively motivated domains of *physical-causal*, *perspectival*, and *social knowledge*.
- Introduced *explicit-implicit* caption contrasts to measure models' ability to match images and captions when the relevant meaning must be inferred from prior events, locations, intended uses, or social context, while holding the naturalistic visual evidence fixed.
- Built stimulus-generation and evaluation pipelines to compare post-hoc aligned unimodal models, contrastive models, and generative vision-language models (VLMs) against human baselines.
- Showed that post-hoc alignment of unimodal models perform poorly on fine-grained image–text matching; contrastive models improve explicit matching but retain inferential gaps; frontier VLMs approach human consensus on some comparisons but still struggle with social-interaction matching and implicit perspectival judgments.

A Computational Framework for Instance-Level Comparison Across Model Representations 2026

- Developed a unified alignment framework using shape analysis to map diverse models into a universal representational space, enabling cross-architecture and cross-modality comparisons of *individual inputs beyond aggregate similarity scores*.
- Identified systematic input properties (e.g. image prototypicality, sentence imageability) that predict representational convergence across vision and language model families.
- Showed that similarity scores from post-hoc aligned unimodal models correlate with human ratings of individual image-text pairs nearly as strongly as CLIP scores.

Shared Semantic Structure in Independently Trained Vision Models and Language Models 2025

- Provided a deeper account of representational alignment between independently trained vision and language models by characterizing its semantic basis, organization across network layers, and relationship to human judgments.
- Used controlled input perturbations to show that cross-modal alignment is driven by deeper semantic content rather than visual appearance: alignment was largely preserved under appearance-only changes but weakened when semantic content was removed.
- Showed that aggregating multiple caption embeddings per image or multiple image embeddings per caption strengthens cross-modal alignment, demonstrating the importance of accounting for many-to-many image-text relationships.

Computational Characterization of Arousal in Adaptive Decision-Making 2024

- Combined Bayesian modeling with pupillometry to characterize how pupil-linked arousal reflects the use of reward and uncertainty in human value-based decision-making under changing reward contingencies.
- Discovered that pupil-linked human arousal system dynamically encodes reward prediction error and estimation uncertainty to guide exploratory behavior in a changing environment.

Research Experience

NeuroML Lab | ***UC San Diego*** Sept. 2024 – Present

Graduate Student Researcher, ***Advisor****: Meenakshi Khosla*

- Investigate shared meaning and representation across vision and language systems through empirical analysis, computational methods, and controlled diagnostic evaluations.

Computational Modeling of Intelligent Systems | ***UCSD/TU Darmstadt*** Sept. 2020 – Sept. 2024

Visiting Scholar/Graduate Student Researcher, ***Advisor****: Angela J. Yu*

- Designed novel behavioral experiments and built mathematical models (Bayesian modeling, Reinforcement Learning) to quantify the computational mechanisms underlying value-based decision-making under uncertainty.

Swartz Center for Computational Neuroscience | ***Institute for Neural Computation*** Jun. – Sept. 2019

Research Intern - Neuroscience for Architecture, ***Co-advisors****: Eduardo Macagno, Tzyy-Ping Jung*

- Authored a 200-page comprehensive lab manual documenting research protocols, biosensing device operation (EEG, eye-tracking, heart-rate monitor), and analysis pipelines to standardize multimodal data collection

Cognitive Development Lab | ***UC San Diego*** Mar. 2018 – Aug. 2020

Research Assistant, ***Advisor****: Gedeon Deák*

- Automated eye-tracking data analysis by developing a tool with a YOLO-based convolutional neural network to classify fixations into regions of interest, eliminating the need for manual labeling.

Selected Talks & Poster Presentations

Representational Alignment Between Vision and Language Models. Cognition at the Shore, UC San Diego, April 29, 2026. (*Research talk*)

Deep Representational Alignment Between Vision and Language Models. Visual Learning Lab, UC San Diego, November 21, 2025. (*Invited guest talk*)

Many-to-Many, Yet Convergent: Insights into the Alignment of Vision and Language Models. Conference on Cognitive Computational Neuroscience (CCN 2025). (*Poster*)

A Pupillometry Study of Reward and Uncertainty in Exploratory Decision Making. European Conference on Visual Perception (ECVP 2023). (*Poster*)

Role of Pupil-Linked Uncertainties and Rewards in Value-Based Decision Making. Conference on Cognitive Computational Neuroscience (CCN 2022). (*Poster*)

Gender Differences in Face-Based Trait Perception and Social Decision Making. 43rd Annual Meeting of the Cognitive Science Society (CogSci 2021). (*Poster*)