Huzheng Yang

[huzeyann.github.io]

Email: huze.yann@gmail.com
Mobile: 215-391-0987

Education

2022 - 2027.05 **University of Pennsylvania**
Computer Science, PhD & Master (GPA: 3.9/4.0), advisor: Jianbo Shi and James Gee

2017 - 2021 **University of Electronic Science and Technology of China**
Computer Science, Bachelor & Mathematics, Bachelor (GPA: 3.8/4.0)

Publications [Google Scholar]

Vibe Spaces for Creatively Connecting and Expressing Visual Concepts *CVPR 2026*
H. Yang*, Katherine Xu, Andrew Lu, Michael D. Grossberg, Yutong Bai, Jianbo Shi*

Artifacts and Attention Sinks: Structured Approximations for Efficient Vision Transformers *arxiv*
Andrew Lu, Wentinn Liao, ***H. Yang****, Jianbo Shi*

Brain Decodes Deep Nets *CVPR 2024 spotlight*
H. Yang*, James Gee, Jianbo Shi*

AlignedCut: Visual Concepts Discovery on Brain-Guided Universal Feature Space *arXiv*
H. Yang*, James Gee, Jianbo Shi*

Memory Encoding Model *Algonauts '23 Comp. Winner*
H. Yang*, James Gee, Jianbo Shi*

Enhancing neural encoding models for naturalistic perception with a multi-level integration of deep neural networks and cortical networks
Yuanning Li,* ***H. Yang**** *, Shi Gu* *Algonauts'21 Comp. Winner*

Interpretable multimodality embedding of cerebral cortex using attention graph network for identifying bipolar
H. Yang* *, Xiaoxiao Li*, Yifan Wu, Siyi Li, Su Lu, James S. Duncan, James C. Gee, Shi Gu* *MICCAI '19*

Research Interests: Computer Vision, GenAI, Multi-modal, Full-stack ML

Open-source package: ncut-pytorch for spectral embedding [website] [GUI software] [slides]

- **Highlights**: multi-modal clustering for VLM/LLM features; correspondence grounding across vision and text; support user-interactive prompt; improved algorithm complexity to $O(n)$, heavy engineering optimization 100x speedup vs. sklearn; analytical solution, training-free, support hot-swapping VLM backbone model; GUI Software CI/CD.
  - **Algorithm**: improved complexity from $O(n^2)$ to $O(n)$, speedup 100x on large graph; improved accuracy 10% by FPS balanced sampling; proposed user-interactive prompt methods that respond in real-time 100ms latency.
  - **Engineering**: built custom memory chunking operations reduce memory complexity to $O(1)$, optimized GPU by efficient caching, optimized CPU by divide-and-conquer. Speedup 10x and support real-time streaming.
  - **Software**: built Gradio GUI software, customized frontend GUI Vue.js; deploy Docker container to HuggingFace, AWS, GCP, and local infrastructure. CI/CD with REST API, served industry and academia collaborators.

AI Competition: large-scale VLM training on infrastructure I built [slides] [paper1][paper2]

- Won The Algonauts Competition(algonauts.csail.mit.edu), build VLM that predicts brain activation.
  - **Infrastructure**: build my own cluster with Docker and NFS, deploy training jobs with Ray Tune, dynamic node join and fault tolerance. Optimized cluster performance with DeepSpeed ZeRO, minimize inter-rack communication.
  - **Algorithm**: surprising method that overtakes 2nd place score by 10% score. We use unexpected combination of multi-modal model input that reveled a bug in the dataset, a new data pre-processing method was developed.

AI Research: understand model mechanism and improve the model

- First understanding how AI models works, then improve the models' machanism (diffusion process, attention layer).
  - **Understand**: probed VLM/LLM models with Algonauts, dissect information flow on attention heads and layers. Found shared visual concepts across supervised and unsupervised vision models. [CVPR spotlight] [video]
  - **Understand**: discovered visual and language concepts in VLM and LLMs, found free-lunch image-text correspondence that help solve visual grounding without supervised training. [arxiv] [slides]
  - **Improve**: discovered attention sink mechanism; proposed Nyström approximation with blocked FPS sampling, speedup VLM/LLM attention layer inference by 10x, compatible with FlashAttention. [arxiv]
  - **Improve**: improved CLIP-conditioned Diffusion model (IP-adapter), compress CLIP feature sparse manifold into a dense manifold, improve sampling plausibility and creativity. [website] [CVPR paper][demo]

Skills

ML Research: VLM/LLMs, Diffusion, VAE, GAN, Recommendation System, RL, MoE, GNN, Gaussian Mixture, NeRF
Full-stack ML: DeepSpeed, Ray Tune, Optuna, ClearML, pytorch-lightning, HuggingFace, OpenCV, Web ONNX
Programming: Python, Pytorch, Tensorflow, Caffe, Gradio, C/C++, CUDA, Rust, Golang, SQL, Shell, JavaScript
Developer Tools: Git, GDB, Perf, MySQL, Redis, MongoDB, WandB, AWS, GCP, JetBrains, Cursor
System & Network Operations: Linux, Docker, Kubernetes, Slurm, Switches, Routers, Physical Servers