CHENGYUAN ZHANG

(425) 753-9631 | cyzhang0339@gmail.com |linkedin.com/in/chengyuan-zhang-0039z/

University of Maryland College Park - Computer Science, Mathematics (2021 - 2025)

RESEARCH

Gamma Lab, UMIACS, UMCP — Undergraduate Research Assistant Fall 2024 - Present

DAVE: Diverse Atomic Visual Elements Dataset

Under review, preprint available on arXiv

- Large-scale video dataset with multi-task annotations (detection, tracking, action localization and recognition) in challenging traffic scenarios.
- Designed and implemented DAVE-DETR, a transformer-based detector variant with a hierarchical query generator and redundancy reduction module, improving detection of small objects in complex environments.
- Benchmarked state-of-the-art detectors as baselines for DAVE, achieved SOTA performance of DAVE-DETR on DAVE (e.g., AP_{50}=0.292, AR_l=0.712), outperforming YOLO and other existing DETR variants like Salience DETR and Deformable DETR (AP_{50}≈0.16-0.25).
- Built a YOLO-based pipeline to automatically blur sensitive information for privacy preservation in data.

Bi-VLM: Ultra-Low Precision PTQ in VLMs

Under review, preprint available on arXiv

- Co-developed Bi-VLM, a post-training quantization method targeting ultra-low-bit (1 bit) for vision-language models (VLMs), improving performance by 4%-45% over SOTA quantizations.
- Implemented ultra-low-bit (1-3 bit) quantization for baselines (AWQ, QVLM) and benchmarked on major VLMs (LLaVA-Next, LLaMA, Qwen) across ScienceQA, MME, MMMU, and VizWiz-VQA, demonstrating Bi-VLM's superior performance.

EXPERIENCES

Z.AI (AI Software Engineer, Intern) **June 2024 - August 2024**

- Optimized Retrieval-Augmented Generation (RAG) by integrating graph RAG and rank RAG, improving top-1 hit rate by 12% on large document collections.
- Implemented keyword and embedding indexing, improving retrieval accuracy across different data.
- Applied adapter/prefix fine-tuning on multimodal GLM4 for video-based reasoning, enhancing detection of dangerous construction behaviors and illegal fishing activities.
- Automated benchmarking process to measure hit rate and answer quality, reducing manual evaluation by 80% and enabling reproducible system comparisons.

Emotion Recognition (BitCamp Hackathon 2023)

- Built a video-based emotion recognition pipeline combining DETR for face detection and a CNN classifier (62% accuracy on FER2013), applied to analyze student engagement in exam settings.

PROJECTS

AI Trip Planner

- Developed and deployed a language-model powered trip planner that generates itineraries from natural language queries, integrating a retrieval-augmented generation (RAG) over Wikivoyage.

- Optimized LangChain's retrieval API by parallelizing embedding calls, reduced retrieval latency and improved efficiency
- Implemented user management with JWT authentication, ensuring security and personalized access for users.

Go Marketplace

- Built a RESTful marketplace backend in Go (Gin, MongoDB) with JWT authentication, product/order management, and a real-time user-to-user live chat system.
- Trained a BERT-based embedding model on Kaggle's Online Sales dataset (masked language modeling) and integrated it into a hybrid recommendation system combining content-based and collaborative filtering for personalized product search.
- Automated API testing with Postman collections and custom test scripts, reducing manual effort by ~70%.

Mini CFO Copilot

- Built a VLM-powered financial analysis agent that answers natural language financial questions on revenue, gross margin, opex, EBITDA, and cash runway, using both textual/numerical data and chart visualizations.
- Designed a multi-step agent involving intent classification, structured JSON extraction (intent, time span, entity), data function calls, and final LLM response based on the functions return.
- Implemented comprehensive unit tests to validate data loading and financial calculations, and integrated a basic RLHF feedback loop collecting user preferences to refine answer style.

PUBLICATIONS/PREPRINTS

- Wang, X., Sandoval-Segura, P., Zhang, C., Huang, J., Guan, T., Xian, R., Liu, F., Chandra, R., Gong, B., & Manocha, D. DAVE: Diverse Atomic Visual Elements Dataset with High Representation of Vulnerable Road Users in Complex and Unpredictable Environments. arXiv preprint arXiv:2412.20042, 2024.
- Wang, X., Huang, J., Abdalla, R., Zhang, C., Xian, R., & Manocha, D. Bi-VLM: Pushing Ultra-Low Precision Post-Training Quantization Boundaries in Vision-Language Models. arXiv preprint arXiv:2509.18763, 2025.