ZHIYUAN QI

Tsinghua University, Department of Electronic Engineering, M.S. in EECS, Beijing, China
+86 13107696517⋄ qzy24@mails.tsinghua.edu.cn or qzy46511786718@gmail.com ⋄ https://togekiss0603.github.io/

RESEARCH INTERESTS

VLM/MLLM, LLM Agent, Multimodal Learning, Spatial Intelligence, World Model, 3D Computer Vision

EDUCATION

Tsinghua University, M.S. in EECS — **GPA:3.73/4.00**, 2024 - present
Beijing JiaoTong University, B.E. in Communication Engineering — **GPA:3.89/4.00**, 2020 - 2024

PUBLICATIONS

1. **Zhiyuan Qi***, Yulong Feng and Zhijin Qin, "Adaptive Sampling and Joint Semantic-Channel Coding Under Dynamic Channel Environment," **ICC 2025** - IEEE International Conference on Communications, Montreal, QC, Canada, 2025, pp. 518-523, doi: 10.1109/ICC52391.2025.11161632.
2. **Zhiyuan Qi***, Jingkai Ying, Yulong Feng, Zhijin Qin, Zhu Han, Rahim Tafazolli and Yonina C Eldar, "Semantic Communication Enabled Holographic Video Processing and Transmission," IEEE Communication Magazine, Accepted (JCR Q1/SCI Q2).
3. **Zhiyuan Qi*(Co-first Author** 1^{st}**)**, "FurniReason: Zero-Shot Vision-Language Reasoning for Manual-Guided Furniture Assembly," **AAAI 2027** (under review), work done while interning at UIUC.
4. Jierui Li*, **Zhiyuan Qi***, Hao Zhu, Yufan Liu, Jixian Liu, Shaojie Jiang, Jianda Wang, Yaqi Liu, Xiaotong Li and Wei Wang **(*Equal contribution)**, "RouteGraph-Mona: Confusion-Aware Routing Fine-Tuning for Mineral Image Classification," **PRICAI 2026**.
5. **Zhiyuan Qi*(Co-first Author** 2^{nd}**)**, "CueShiftBench: How Heuristic Cues Shape Multimodal Reward Evaluation," **NAACL 2027** (under review).

RESEARCH EXPERIENCE

FurniReason: Zero-Shot Vision-Language Reasoning for Manual-Guided Furniture Assembly. 2026.3 - 2026.5
Research Intern | University of Illinois Urbana-Champaign (**UIUC**) | Supervisor: Heng Ji | Champaign, IL in US

- Developed **FurniReason**, a **zero-shot** vision-language reasoning framework for assembling furniture from component configurations using only IKEA-style assembly manuals and MLLMs.
- Proposed an **anchor-guided manual-to-operation** method that generates part-level instructions from page-level manuals aligned with intermediate assembly states, resulting in a hierarchical assembly graph.
- Designed a **VLM-only 6D pose-solving agent** that iteratively updates the 6D pose of each part to assemble furniture in real-world 3D space through a rendering–judging–proposing–dreaming–checking–refining workflow. Based on the assembly manuals, current state, assembly graph, and operation memory, the agent generates a sequence of translations and rotations and performs 6D pose upating for each part, enabling step-by-step assembly in the space.
- Improved assembly-graph accuracy from 28.24% to 53.92% over the baseline and enhanced local contact and geometric alignment in reference-free 6D pose estimation. Further analyzed failure modes, revealing both the potential and limitations of general-purpose MLLMs for zero-shot furniture assembly.

CueShiftBench: How Heuristic Cues Shape Multimodal Reward Evaluation. 2026.2 - 2026.7
Research Intern | Tsinghua University (**THU**) | Supervisor: Zhiyuan Liu | Beijing, China

- Introduced **CueShiftBench**, a controlled diagnostic **benchmark for multimodal reward models** and **MLLM-as-a-Judge** evaluators, designed to examine whether MLLMs rank responses based on visual evidence and task requirements or rely on superficial heuristic cues.
- Constructed 1,200 minimal-difference response pairs across three visual domains including charts, VQA scenes and documents to systematically isolate the effects of individual heuristic cues.
- Designed a bidirectional **A/B–B/A evaluation protocol** that decomposes evaluation outcomes into stable preference for the criterion-based response.

- The Full condition increased the selection rate of image-correcting responses by 45.2% relative to the No-image condition. High-strength redundant elaboration reduced the corresponding preference by 97.4%, while authoritative wording and moderate-strength quality claims increased it by 46.2% and 70.7%.

RouteGraph-Mona: Confusion-Aware Routing Fine-Tuning for Mineral Image Classification. 2026.3 - 2026.6
Student Researcher |

- Developed **RouteGraph-Mona**, a lightweight route-space regularization method built on the parameter-efficient adapter for mineral image classification, addressing substantial intra-class appearance variations and high inter-class visual similarity.
- Proposed **class-wise route anchors** and **confusion graph** based on regularized routing signatures from adapter parameters tuning. The route anchors encourage class-consistent routing patterns, while the confusion-weighted margins promote greater separation between visually similar mineral categories in the routing space. The resulting branch-selection behavior defines a compact routing space that characterizes the scale preferences of individual images.
- On the **MinetV2** dataset, the full model reduced total off-diagonal confusion errors by 15.3% and increased the routing-space separation between confusing class pairs to approximately 1.90 times its original value.

A Holographic Classroom Based on Multi-View 3D Reconstruction and Digital Avatar 2025 - present
Student Researcher | Tsinghua University, Department of Electronic Engineering | Beijing, China

- Develop a semantic-driven sampling and compression algorithm for 3D point clouds, which significantly reduces data volume while preserving key information(such as facial features and details), thereby enhancing transmission real-time performance and quality.
- Improve 3D human pose estimation framework and optimize real-time digital human generation through asynchronous pipelining.
- The teacher's point cloud is captured via 4 Microsoft Azure Kinect DK camera arrays and reconstructed through fusion algorithm. The student's digital avatar is driven by key facial and limb parameters captured through a monocular RGB camera.
- 10 fps reconstruction frame rate at 200k points/frame scale (200ms latency). Student-side digital human feature stream consumes only 20KB/s bandwidth (100ms latency).

Image Transmission with Joint Sampling and Coding at Adaptive Channel Environments 2024 - 2025
Student Researcher | Tsinghua University, Department of Electronic Engineering | Beijing, China

- Propose a semantic-aware sampling and reconstruction algorithm, dynamically optimize the number of samples in each region of an image. The sampling matrix for each region is individually optimized based on semantic importance, yielding a semantic sampling proportion distribution map shared with the receiver.
- Propose a semantic distribution map-assisted reconstruction algorithm to achieve high fidelity.
- Design an attention-based channel adaptation module to overcome the mismatch between neural network models trained under training channel conditions and those tested under actual channel conditions.
- Achieve reconstruction performance nearly equivalent to that of full data at a 30% sampling rate.
- Noise robustness is enhanced, achieving over 50% sampling rate reduction across diverse channel conditions, outperforming other models across at least a 20dB range of environmental SNR variations.

AI Psychological Counselor Based on Knowledge Graphs and Emotion Analysis 2022 - 2023
Student Researcher | Beijing JiaoTong University, Department of Electronic Engineering | Beijing, China

- For the professional psychology Q&A module, knowledge graph-driven procedure is achieved through intent recognition through LR+GBDT multi-model fusion and Bert+TextCNN, coupled with entity recognition via BiLSTM-CRF+AC automata. Response quality is enhanced through RL polling strategies.
- For the counselor script generation module, finetune UniLM on targeted datasets across different scenario categories. Based on an enhanced Encoder architecture, it employs a specific self-attention mask mechanism to achieve seq2seq modeling, generating targeted and comforting psychological counseling scripts.

Semantic Communication Enabled Holographic Video Processing and Transmission 2025 - 2026
Student Researcher | Tsinghua University, Department of Electronic Engineering | Beijing, China

- Proposed a semantic-aware architecture for **holographic video communication (HVC)**, investigating 3D data representations and system requirements and examining the role of semantic communication in enabling efficient HVC systems.

- Developed a semantic-driven joint optimization method for point-cloud sampling-semantic-channel coding, downsampling point clouds according to their local and global semantic features, improving communication efficiency while preserving the quality of the transmitted point clouds. At sampling rates below 12.5%, it improved point-cloud recognition performance on **ModelNet40** by 92.98% over **PointNet++**.

JOB EXPERIENCE

Baidu Inc. | ACG Group 2026.7 - present
Multimodal Large Language Model (MLLM) Algorithm Engineer | Beijing, China

- Developed an integrated multimodal event detection system for transportation and government-service scenarios through MLLM-expert model fusion. The UniVL foundation model, inspired by the InternVL2.5 architecture, consists of a 6B visual feature extraction backbone, a modality-adaptive MLP connector, and a 30B InternLM decoder. It was pretrained on hundreds of millions of image–caption, object detection, and OCR–image-text pairs, followed by SFT on VQA, classification, and over 1M industry-specific samples from government, transportation, and industrial domains to improve domain understanding and cross-scenario generalization.
- For highway surveillance videos, used an industry-specific RT-DETR detector to localize pedestrians, construction workers, and vehicles and extract $[N, 192]$ visual features. Projected these features into the LLM latent space through a lightweight 4M-parameter MLP projector, encoded the corresponding bounding boxes and class labels as structured text, and performed LoRA fine-tuning on UniVL-8B. Improved the detection of traffic congestion, construction workers, and pedestrian intrusions, achieving P/R of 0.9735/0.9649, 0.9536/0.9664, and 0.8761/0.8839.
- Proposed a layer-wise feature fusion strategy that combines visual tokens extracted from the UniVL-8B vision tower at 25%, 50%, and 75% depths with features from the domain-specific RT-DETR traffic-scene expert detector, and injects the fused features into the 0th, 1st, and 2nd layers of the LLM decoder, respectively. Achieved P/R of 0.9820/0.9735, 0.9730/0.9664, and 0.8609/0.8839.

SELECTED HONORS AND AWARDS

National Scholarship (Awarded to the **top0.2**% of students **nationwide**) 2022
Outstanding Undergraduate Thesis Award 2024
Tsinghua University Comprehensive **1st-Class** Scholarship 2025
The China Graduate Electronic Design Contest **2nd** Prize 2025
MCM (Mathematical Contest In Modeling)/ ICM (Interdisciplinary Contest In Modeling), **Finalist** Prize (Ranked in the **top1**% of teams **worldwide**) 2021

PATENTS

1. "A Channel Environment Adaptive Joint Optimization Method and System for Sampling-Semantic-Channel Coding", ZL 2024 1 1509807.8, **Granted**.
2. "A Joint Optimization Method, Device, and Medium for Sampling-Semantic in Point Clouds for Holographic Communications", ZL 2025 1 1141478.0, **Granted**.
3. "A Method for Assessing the Quality of Experience of 3D Representations in Immersive Communications," Patent pending.
4. "A Holographic Classroom System for Generating Digital Humans via Multi-View Point-Cloud Reconstruction and Monocular-Camera Human Pose Estimation," Software copyright registered.

RELATED COURSES

3D Computer Vision (4.0/4.0) | Convex Optimization (4.0/4.0) | Intellitent Wireless Communication (4.0/4.0)
Calculus (4.0/4.0) | Linear Algebra (4.0/4.0) | Probability Theory and Mathematical Statistics (4.0/4.0)
Data Structure (4.0/4.0) | Signal and System (4.0/4.0) | Digital Signal Processing (4.0/4.0)

SKILLS

Python, Pytorch, MATLAB, C/C++, Unity, Isaaclab, Microsoft Azure Kinect DK, Git, Docker, Conda, CMake, Linux

SELECTED SERVICES

Reviewer for conferences and journals:

- **2026**: CVPR, IEEE VTC Spring, NeurIPS, PRICAI
- **2025**: IEEE ICME, IEEE COMMAG