



Zhe Cao

📍 Nanjing, Jiangsu • 📞 (+86)176-0149-7616 • ✉️
zhcao181@gmail.com



🎓 EDUCATION

Nanjing University
B.S. in Computer Science and Technology

Sep. 2023 – Present

- GPA: **4.10/5.0**; core coursework includes Data Structures and Operating Systems.

⚙️ SKILLS

- **Programming and Frameworks:** Python, C++, Shell; PyTorch, Transformers, VeRL, RLinf
- **LLM Training and Inference:** LoRA fine-tuning, PPO/GRPO reinforcement learning, FP8 mixed-precision training, multi-GPU training
- **Engineering and English:** Linux, Docker, Git; CET-4: 567/710, CET-6: 505/710; proficient in reading and writing English research papers and technical documentation

💼 INTERNSHIP EXPERIENCE

Insta360
Multimodal Algorithm Intern

Apr. 2026 – July. 2026

- **Visual Understanding and Camera Parameter Recommendation**
 - **Project Context:** Developed an intelligent shooting assistant for novice photographers who struggle to configure camera parameters based on the current scene and their shooting intent.
 - **Technical Approach:** Fine-tuned a VLM to better understand scene content, lighting, subject placement, and natural-language intent. Constructed realistic user intents from collected user data, enabling the model to recommend exposure, white balance, focal length, and composition settings directly from user requests.

AgiBot
Inference Deployment Intern

Jan. 2026 – Mar. 2026

- **Project Context:** Optimized training and inference pipelines for deploying embodied models on edge-computing platforms and supported sim-to-real transfer of reinforcement-learning policies to physical robots.
- **On-Device Inference Acceleration:** Built an FP8 mixed-precision training and inference validation pipeline with **Transformer Engine**; compared BF16 and FP8 in accuracy, memory use, throughput, and latency to support deployment decisions.
- **Reinforcement Learning Deployment:** Contributed to real-robot policy deployment with the **RLinf** framework and supported validation of policy transfer from simulation to physical platforms.

🔬 RESEARCH AND PROJECT EXPERIENCE

Link Lab, Nanjing University
Research Direction: Omni-modal Video Intelligence and Evaluation

Jul. 2025 – Jun. 2026

- **Research Focus:** Studied the joint modeling of visual actions, speech, music, and environmental sounds through omni-modal data construction, model post-training, long-video understanding, generative evaluation, and hallucination reliability, forming a continuous research program from generation evaluation to model training, long-video understanding, and reliability calibration.
- **T2AV-Compass: A Unified Benchmark for Text-to-Audio-Video Generation**
 - **Technical Approach:** Introduced a unified T2AV benchmark with **500** complex prompts, using semantic clustering and video inversion to construct a diverse, high-coverage evaluation set. Combined conventional automatic metrics with an MLLM judge to assess generation quality, temporal consistency, and semantic alignment.
 - **Project Deliverables:** Designed and implemented evaluation pipelines for cross-modal synchronization, audio instruction following, and video instruction following; completed unified inference, visualization, and fine-grained capability analysis across multiple generative models.
 - **Publication:** Published at **ICML 2026 (first author)** and selected for an **Oral at the CVPR 2026 VGBE Workshop**; <https://arxiv.org/abs/2512.21094>.
- **OmniVideoBench: A Benchmark for Audio-Visual Collaborative Reasoning**
 - **Data and Evaluation:** Contributed to a benchmark of **628** real-world videos and **1,000** high-quality QA pairs spanning **13** reasoning categories, each accompanied by a step-by-step reasoning chain.
 - **Project Deliverables:** Conducted unified evaluation and error analysis across open- and closed-source MLLMs, revealing substantial human-model gaps in cross-modal evidence binding, long-video reasoning, and audio-visual understanding.
 - **Publication:** Published at **ICLR 2026**; <https://arxiv.org/abs/2510.10689>.
- **AVSCap: Orchestrating Audio-Visual Synergy for Omni-modal Video Captioning**
 - **Data and Post-Training:** Contributed to **AVSCap-130K**, a tri-modal corpus built with a decoupled-then-fused annotation and automatic-verification pipeline. Trained AVSCap-7B with a two-stage SFT + GRPO strategy and hybrid rewards targeting acoustic completeness and audio-visual synergy event recall.
 - **Evaluation and Project Deliverables:** Contributed to AVSCapBench, comprising **1,226** clips of 30–120 seconds. AVSCap-7B achieved **60.44** Total Recall, **82.4** UGC-VidCap Avg., and **66.6** on Daily-Omni; experiments showed that reinforcement learning yielded larger gains than simply scaling SFT data.

- **Publication:** EMNLP 2026; <https://arxiv.org/abs/2607.12820>.
- **OmniHalluc-L: Counterfactual Benchmarking and Reliability Calibration for Long-Form Omni Hallucination**
 - **Evaluation Framework:** Proposed a counterfactual event-binding protocol that constructs supported/counterfactual claim pairs from the same audio-visual evidence and uses strict-pair accuracy to evaluate speaker, temporal, and modality binding. The benchmark contains **3,600** single-claim QA items from **638** long videos totaling **256.87** hours.
 - **Method and Project Deliverables:** Proposed the frozen-backbone MPRC reliability-calibration method, combining responses to negative audio perturbations with native audio-visual confidence. It improved pair accuracy from **32.06%** / **41.55%** to **36.22%** / **51.09%** for Qwen2.5-Omni-7B / Qwen3-Omni-Instruct.
 - **Publication:** EMNLP 2026; <https://arxiv.org/abs/2606.03614>.

Tsinghua University NLP Group Multi-Agent Systems

2025.09 ~ 2026.01

- **Research Focus** Focused on multi-agent systems, multi-model collaboration for Deep Research, and the development of corresponding reward models.

Reasoning and Learning Group, Nanjing University

Aug. 2024 – Mar. 2025

Research Direction: LLM Safety and Trustworthiness Evaluation

- **Research Focus:** Studied safety consistency, jailbreak robustness, and fine-grained risk evaluation in multi-turn LLM conversations, with emphasis on safety degradation across languages, conversational depth, and attack strategies.
- **SafeDialBench: A Multi-Turn Dialogue Safety Benchmark**
 - **Data and Evaluation:** Co-designed an evaluation framework covering **6** safety dimensions and **22** dialogue scenarios, constructed **4,000+** Chinese and English multi-turn dialogue instances, and incorporated diverse jailbreak strategies into a fine-grained safety test set.
 - **Project Deliverables:** Evaluated **17** leading LLMs and analyzed safety-consistency degradation across languages, dialogue turns, and attack strategies, producing a reusable protocol for multi-turn safety alignment and defense research.
 - **Publication:** Published at **ICLR 2026**; <https://arxiv.org/abs/2502.11090>.

Awesome Latent Space

Jan. 2026 – Present

Open-Source Research Project

- **Research:** Contributed to writing and maintaining an open-source survey of latent-space research, organizing the field around foundations, evolution, mechanisms, capabilities, applications, and future directions.
- **Project Deliverables:** The repository has received **900+ GitHub Stars** and ranked second on Hugging Face Daily Papers; <https://github.com/YU-deep/Awesome-Latent-Space>.
- **Publication:** <https://arxiv.org/abs/2604.02029>.

 **CAMPUS ACTIVITIES**

Nanjing University Tencent Rhino-Bird Club

Sep. 2024 – Jun. 2026

Technology Innovation Department Member

- Organized technical seminars, research-tool workshops, and academic blog writing to promote technical exchange and research knowledge sharing across campus.