Gengyuan Zhang

Oettingenstr. 67 EU/103, 80538, Munich, Germany

gengyuanmax@gmail.com / zhang@dbs.ifi.lmu.de   LinkedIn
https://gengyuanmax.github.io/   Google Scholar   Github

Experience

Oct. 2021 – Present

Research Assistant, Department of Informatics, Ludwig Maximilian University of Munich (LMU), Munich, Germany
Conducting research on multimodal learning and video understanding
Mentoring and supervising master/bachelor thesis
Assisting and coordinating teaching responsibilities:

- **Tutorial**: Machine Learning (since SS2022)
- **Advanced Seminar**: Machine Learning with Knowledge Graphs, Foundation Models in AI
- **Practical Course**: Connecting Language to Vision

Mar. 2020 – Nov. 2020

Intern, Agile Robots AG, Munich, Germany
Developed an automated hand-to-eye camera calibration pipeline
Deployed and tested a robotic object localization and grasping system

Sept. 2019 – Feb. 2020

Student Assistant, Department of Informatics, Technical University of Munich (TUM), Munich, Germany
Designed and implemented a perception stack in RobMoSys (funded by European Horizon 2020)
Developed computer vision algorithm on robots, including object detection and recognition modules

Education

Oct. 2021 – Present

Ph.D. Computer Science, Ludwig Maximilian University of Munich (LMU), Munich, Germany
Advisor: Prof. Dr. Volker Tresp
Thesis (provisional): Multi-event Video Understanding and Reasoning

Oct. 2018 – Jul. 2021

M.Sc. Electrical Engineering and Information Technology, Technical University of Munich (TUM), Munich, Germany
Grade: 1.3/1.0

Sept. 2014 – Jul. 2018

B.Eng. Opto-Electronics Information Science and Engineering, Zhejiang University, Hangzhou, China
Grade: 3.73/4.00

Research Publications

Conference Proceedings

1. **G. Zhang**, J. Ren, J. Gu, and V. Tresp, “Multi-event video-text retrieval,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023)*, 2023, pp. 22 113–22 123.
2. R. Amoroso*, **G. Zhang***, R. Koner, L. Baraldi, R. Cucchiara, V. Tresp, *et al.*, “Perceive, query & reason: Enhancing video qa with question-guided temporal queries,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2025)*, *Equal contribution, 2025.

3. **G. Zhang**, Y. Zhang, K. Zhang, and V. Tresp, "Can vision-language models be a good guesser? exploring vlms for times and location reasoning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2024)*, 2024.
4. Y. Zhang, H. Chen, A. Frikha, Y. Yang, D. Krompass, **G. Zhang**, J. Gu, and V. Tresp, "CL-CrossVQA: A continual learning benchmark for cross-domain visual question answering," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2025)*, 2025. URL: https://arxiv.org/abs/2211.10567.
5. R. Liao, M. Erler, H. Wang, G. Zhai, **G. Zhang**, Y. Ma, and V. Tresp, "VideoINSTA: Zero-shot long video understanding via informative spatial-temporal reasoning with LLMs," in *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2024)*, 2024.

Preprints

1. **G. Zhang***, M. L. A. Fok*, Y. Xia, Y. Tang, D. Cremers, P. Torr, V. Tresp, and J. Gu, "Localizing events in videos with multimodal queries," *arXiv preprint*, vol. arXiv:2406.10079, Jun. 2024, *Equal contribution. URL: https://arxiv.org/abs/2406.10079.
2. **G. Zhang**, M. Ding, T. Liu, Y. Zhang, and V. Tresp, "Memory helps, but confabulation misleads: Understanding streaming events in videos with mllms," *arXiv preprint*, vol. arXiv:2502.15457, 2025. URL: https://arxiv.org/abs/2502.15457.
3. **G. Zhang***, J. Bi*, J. Gu, Y. Chen, and V. Tresp, "SPOT! revisiting video-language models for event understanding," *arXiv preprint*, vol. arXiv:2311.12919, 2023, *Equal contribution. URL: https://arxiv.org/abs/2311.12919.
4. T. Liu, Z. Lai, **G. Zhang**, P. Torr, V. Demberg, V. Tresp, and J. Gu, "Multimodal pragmatic jailbreak on text-to-image models," *arXiv preprint*, vol. arXiv:2409.19149, 2024. URL: https://arxiv.org/abs/2409.19149.
5. H. Chen, H. Li, Y. Zhang, **G. Zhang**, J. Bi, P. Torr, J. Gu, D. Krompass, and V. Tresp, "FedBiP: Heterogeneous one-shot federated learning with personalized latent diffusion models," *arXiv preprint*, vol. arXiv:2410.04810, 2024. URL: https://arxiv.org/abs/2410.04810.
6. J. Gu, Z. Han, S. Chen, A. Beirami, B. He, **G. Zhang**, R. Liao, Y. Qin, V. Tresp, and P. Torr, "A systematic survey of prompt engineering on vision-language foundation models," *arXiv preprint*, vol. arXiv:2307.12980, 2023. URL: https://arxiv.org/abs/2307.12980.

Skills

Languages	Chinese (native speaker), English (proficient), German (intermediate)
Coding	Python, LaTeX, Matlab, C++

Miscellaneous Experience

Academic Service

- **Conference Reviewer**: CVPR2024-2025, ECCV2024, ICCV2025, NeurIPS2024, COLING2024, ARR2025