

Gengyuan Zhang

Research on multimodal intelligence | Munich, Germany

✉ gengyuanmax@gmail.com

🌐 gengyuanmax.github.io

🌐 LinkedIn

🎓 Google Scholar

🐙 GitHub

Education

- Oct. 2021 – Present  **Ph.D. in Computer Science, Ludwig Maximilian University of Munich (LMU), Munich, Germany**
Expected completion: December 2026
Supervisor: Prof. Dr. Volker Tresp
Thesis: *Promises and Limits of Language-centered Visual Understanding*
- Oct. 2018 – Jul. 2021  **M.Sc. in Electrical Engineering and Information Technology, Technical University of Munich (TUM), Munich, Germany**
Grade: 1.3 (German grading scale; 1.0 is the highest grade)
- Sept. 2014 – Jul. 2018  **B.Eng. in Information Science and Engineering, Zhejiang University, Hangzhou, China**
Grade: 3.73/4.00

Experience

- Oct. 2021 – Present  **Research Assistant, Department of Informatics, Ludwig Maximilian University of Munich (LMU), Munich, Germany**
Research on multimodal learning, visual/video understanding, and multi-modal agents.
Developed large-scale multimodal datasets and evaluation suites for video-language models.
Team lead for a cross-institute research project on intelligent agents for surgical video understanding.
Served as a tutor and course lead and supervised master's and bachelor's theses
- Mar. 2026 – Jul. 2026  **Visiting PhD Student, Department of Engineering, University of Oxford, Oxford, United Kingdom**
Visiting researcher in the Torr Vision Group (TVG) at University of Oxford.
Built an agentic environment for decision-making in wargames.
Designed, implemented, and evaluated memory mechanisms for agentic decision-making under uncertainty in wargame environments.
- Apr. 2025 – Oct. 2025  **Applied Scientist Intern, Amazon, London, United Kingdom**
Built omni-modal audio–video retrieval systems, including end-to-end and multi-embedding architectures.
Evaluated and analysed retrieval robustness across modality-specific query scenarios.
- Mar. 2020 – Nov. 2020  **Intern, Agile Robots AG, Munich, Germany**
Implemented perception models for robotic object localization and grasping.
Developed real-time perception modules for industrial robot systems.
- Sept. 2019 – Feb. 2020  **Student Assistant, Department of Informatics, Technical University of Munich (TUM), Munich, Germany**
Designed and implemented the perception stack for RobMoSys robotic platform, funded by EU Horizon 2020.

Research and Technical Skills

Multimodal ML	📌 Multimodal learning; video understanding; audio–video models.
Agent Systems	📌 Agentic environments; memory mechanisms; agentic tooling.
RL	📌 Reinforcement learning with GRPO; Agentic training.
Evaluation	📌 Benchmark development; retrieval robustness; memory mechanism evaluation.
Frameworks	📌 PyTorch, Transformers, verl, vLLM.
Programming	📌 Python, C++, MATLAB, $\LaTeX$
Languages	📌 Chinese (native), English (proficient), German (intermediate)

Funding, Service and Talks

Funding	📌 Secured EUR 130,000 in funding from BMFTR (German Federal Ministry of Research, Technology and Space) for the individual research project <i>Towards an Intelligent Agent for Streaming Surgical Video Understanding</i> , under Software Campus 2025.
Reviewing	📌 CVPR 2024–2026; ECCV 2024–2026; NeurIPS 2024–2026; COLING 2024; ARR 2025–2026.
Talks	📌 Contributed talk at the 2nd Workshop on Responsible Generative AI at CVPR 2025 (Best Paper Award). Invited talk on “Surgical Video Understanding Agents” at Zeiss Research Center.

Publications and Preprints

* Equal contribution.

- 1 **G. Zhang**, X. Han, S. Zhou, A. Koukash, G. Ghazaei, J. Gu, J. Wu, Y. Zhu, N. Navab, and V. Tresp, “How to Train a Surgeon? Benchmarking Generalist Agents in Surgical Scene Understanding,” in *NeurIPS 2026*, Accepted (poster), 2026. [link](#).
- 2 **G. Zhang**, X. Han, X. Xie, T. Liu, and V. Tresp, “Rethinking Multi-Image Re-Representation in Multi-Image Understanding,” *arXiv:2609.39363*, 2026. [link](#).
- 3 **G. Zhang**, M. Ding, J. Wu, R. Liao, and V. Tresp, “ReEXplore: Improving MLLMs for Embodied Exploration with Contextualized Retrospective Experience Replay,” in *EMR Workshop at ECCV 2026*, 2026. [link](#).
- 4 **G. Zhang***, T. Hannan*, H. Kleiner, B. Aydemir, X. Xie, J. Lan, T. Seidl, V. Tresp, and J. Gu, “AViLA: Asynchronous Vision-Language Agent for Streaming Multimodal Data Interaction,” *arXiv:2506.18472*, 2025. [link](#).
- 5 **G. Zhang***, M. L. A. Fok*, J. Ma, Y. Xia, D. Cremers, P. Torr, V. Tresp, and J. Gu, “Localizing Events in Videos with Multimodal Queries,” in *CVPR 2025*, 2025. [link](#).
- 6 R. Amoroso*, **G. Zhang***, R. Koner, L. Baraldi, R. Cucchiara, and V. Tresp, “Perceive, Query & Reason: Enhancing Video QA with Question-Guided Temporal Queries,” in *WACV 2025*, 2025. [link](#).
- 7 **G. Zhang***, M. Ding*, T. Liu, Y. Zhang, and V. Tresp, “Memory Helps, but Confabulation Misleads: Understanding Streaming Events in Videos with MLLMs,” in *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*, 2025. [link](#).
- 8 **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 *WACV 2024*, 2024. [link](#).
- 9 **G. Zhang***, J. Bi*, J. Gu, Y. Chen, and V. Tresp, “SPOT! Revisiting Video-Language Models for Event Understanding,” *arXiv:2311.12919*, 2023. [link](#).
- 10 **G. Zhang**, J. Ren, J. Gu, and V. Tresp, “Multi-Event Video-Text Retrieval,” in *ICCV 2023*, 2023. [link](#).
- 11 Z. Han*, **G. Zhang***, Y. Ma, and V. Tresp, “Time-dependent Entity Embedding is not All You Need: A Re-evaluation of Temporal Knowledge Graph Completion Models under a Unified Framework,” in *EMNLP 2021*, 2021. [link](#).
- 12 Y. Zhu, J. Yu, L. Li, **G. Zhang**, Z. Yang, S. Staab, P. Dokania, P. Torr, J. Tang, and E. Kharlamov, “Conformalized Large Language Models under Configuration Shift,” in *EMNLP 2026 (main conference)*, Accepted, 2026. [link](#).
- 13 H. Jiang, J. Ma, G. Du, J. Sun, **G. Zhang**, Z. Zhang, and K. Luo, “Approximate or Perish: Spectral MLP-KAN Diffusion with Attentive Function Learning for Unsupervised Hyperspectral Image Restoration,” in *CVPR 2026 Workshops (PBVS)*, 2026. [link](#).

- 14 T. Liu, Z. Lai, J. Wang, **G. Zhang**, S. Chen, P. Torr, V. Demberg, V. Tresp, and J. Gu, "Multimodal Pragmatic Jailbreak on Text-to-image Models," in *ACL 2025 (Long Papers)*, 2025. [link](#).
- 15 J. Lan, U. Schlegel, T. Hannan, **G. Zhang**, H. Chen, and T. Seidl, "Unveiling the "Fairness Seesaw": Discovering and Mitigating Gender and Race Bias in Vision-Language Models," *arXiv:2505.23798*, 2025, Revised February 2026. [link](#).
- 16 H. Chen, H. Li, Y. Zhang, J. Bi, **G. Zhang**, Y. Zhang, P. Torr, J. Gu, D. Krompass, and V. Tresp, "FedBiP: Heterogeneous One-Shot Federated Learning with Personalized Latent Diffusion Models," in *CVPR 2025*, 2025. [link](#).
- 17 Y. Zhang, H. Chen, A. Frikha, D. Krompass, **G. Zhang**, J. Gu, and V. Tresp, "CL-Cross VQA: A Continual Learning Benchmark for Cross-Domain Visual Question Answering," in *WACV 2025*, 2025. [link](#).
- 18 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 *Findings of EMNLP 2024*, 2024. [link](#).
- 19 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:2307.12980*, 2023. [link](#).