Yan Ma

860+ citations, 1.3k+ GitHub stars across research projects

(+86) 152-6982-9096 • yanma23@m.fudan.edu.cn • Homepage • Google Scholar • Github • Research Statement

Research Interests

My research aims to build multimodal agents for partially observed environments, especially how they perceive, reason, act, and predict over long-horizon tasks. So far, my work has mainly studied how RL changes multimodal models, including training dynamics, the joint improvement of reasoning and perception, and agentic vision tool use. Recently, I am interested in predictive multimodal learning from video, especially how video data can be cleaned, filtered, and annotated to better support the anticipation of future observations and action outcomes.

Education

- **Ph.D. Student, Fudan University** (Advisors: Prof. Pengfei Liu and Prof. Yu Qiao) *2023.09 – 2027.06*
- **Master in Computer Application Technology, Fudan University** *2020.09 – 2023.06*
- **Bachelor in Computer Science, Dalian University of Technology** (GPA: 4.076/5) *2016.09 – 2020.06*

Experience

- **MiniMax & HailuoAI** *(2025.01 – 2026.02)*: worked on multimodal foundation models for visual reasoning and vision tool use in the M-series, and image caption systems for Hailuo video generation
- **Shanghai AI Laboratory** *(2023.07 – 2025.01)*: worked on text generation and unified multimodal model training

Publications

- **VidaForge: Building a Video Foundation Model Pretraining Data Pipeline from Scratch in an Academic Lab** [🌐][GitHub]. *Technical Blog*, July 4, 2026

  Yan Ma, Jiadi Su, Zhulin Hu, Ethan Chern, Linhao Zhang, TianTian Mi, Pengfei Liu

  TLDR: We document a practical pipeline for building video foundation model pretraining data from scratch.

- **What Does Vision Tool-Use Reinforcement Learning Really Learn? Disentangling Tool-Induced and Intrinsic Effects for Crop-and-Zoom** [PDF][GitHub]. *ICML 2026*

  Yan Ma†, Weiyu Zhang†, Tianle Li, Linge Du, Xuyang Shen, Pengfei Liu

  TLDR: We show current vision tool-use RL gains are driven mainly by intrinsic capability improvement and reduced tool-induced harm, not true tool-based correction.

- **One RL to See Them All: Visual Triple Unified Reinforcement Learning** [PDF][GitHub]. *arXiv:2505.18129* [COLM 2026 under review]

  Yan Ma†, Linge Du†, Xuyang Shen†, Shaoxiang Chen, Pengfei Li, Qibing Ren, Lizhuang Ma, Yuchao Dai, Pengfei Liu, Junjie Yan

  TLDR: We build a unified RL system (V-Triune) for jointly training visual reasoning and perception in VLMs, yielding broad gains across both task types.

- **Rethinking RL Scaling for Vision Language Models: A Transparent, From-Scratch Framework and Comprehensive Evaluation Scheme** [PDF][GitHub]. *arXiv:2504.02587*

  Yan Ma, Steffi Chern, Xuyang Shen, Yiran Zhong, Pengfei Liu

  TLDR: We introduce a transparent, from-scratch RL framework and trajectory-level evaluation for VLMs, showing stronger generalization than SFT and clearer training-dynamics insights.

- **ANOLE: An Open, Autoregressive, Native Large Multimodal Models for Interleaved Image-Text Generation** [PDF][🌐][GitHub]. *arXiv:2407.06135; extended version accepted to ICLR 2025 Blog Track*

  Ethan Chern*, Jiadi Su*, Yan Ma*, Pengfei Liu

  TLDR: We present an open native autoregressive multimodal model for coherent interleaved image-text generation, with efficient fine-tuning from Chameleon.

- **MoPS: Modular Story Premise Synthesis for Open-Ended Automatic Story Generation** [PDF][GitHub]. *ACL 2024*

  Yan Ma, Yu Qiao, Pengfei Liu

Technical Skills & Academic Service

- Verl, Transformers, Ray, FSDP, vLLM, SGLang, Jax, FFmpeg, Codex, Git, NeoVim, Tmux
- Reviewer for AAAI 2025, NeurIPS 2025, COLM 2025–2026, ICLR 2026, ICML 2026, and ACL ARR 2026