Ye Li

Email: lye11965@gmail.com
Homepage: memoryoffish.github.io GitHub: github.com/memoryoffish

Seeking: research internship in inference acceleration (LLM / video generation) and code world models.

Education

- **Zhejiang University** Incoming, 09/2027
  M.Eng. in Electronic Information · ZIP Lab · Advisor: Prof. Bohan Zhuang
- **Zhejiang University** 09/2023 – 06/2027
  B.Eng. in Software Engineering, College of Computer Science and Technology GPA: 92.41/100

Research & Projects

Video Inference Infrastructure

- **Inferix 2.0: Inference Framework for Video Generation and World Models** 06/2026 – Present
  Framework core abstraction (inferix_core)
  - Autoregressive video generation: block-causal inference, rolling KV cache, speculative decoding.
  - Pluggable attention, KV cache, external memory and quantization; new models plug in through recipes.
  - Coding agents search speedup combinations over these interfaces: **7.0**× on Wan2.1-1.3B at 720p.

Video & Multimodal Generation

Post-Training

- **Gen-RSI: Recursive Self-Improvement of Video Generators through World Proxies** ICLR 2027
  Second student author · post-training implementation and experiments *Under review · avg. score 3.33/4*
  - Rebuilds the student's videos as executable Blender proxies and repairs them by physics simulation.
  - Repaired renders condition a frozen VACE teacher for on-policy distillation, over 4 self-improvement rounds.
  - Wan2.1-14B: PhyGenBench **0.369**→**0.558**, VBench-2.0 physics **73.09%**→**80.00%**.

Code-Driven World Models

- **Any Time, Any Space: A World-State-Centric Framework for Agentic Video Generation** ICLR 2027
  Co-first author (equal contribution) · code, experiments and demo *Under review · avg. score 3/4*
  - 3D point-cloud world state plus an explicit event chain for time; #2 on Hugging Face Daily Papers.
  - Scene memory of registered static geometry and event memory indexed by scene, event and time.
  - ST-Bench **+7.6–19.4%** over the strongest baseline on four generators; first on all six human-eval criteria.
- **ReCA: Multi-Shot Long Video Extrapolation via Recursive Context Allocation** arXiv 2026
  Second student author · recursive planning and context allocation *NeurIPS 2026 reviews: Weak Accept*
  - Introduced the multi-shot video extrapolation task and the MSVE-Bench / NB-Q benchmarks.
  - Recursive split into context-bounded subproblems, with a frozen short-video model at the leaves.
  - **8–16%** above the strongest baseline, consistency **+28–43%**, **3–4**× faster generation.

Open Source

- **Film-Director-plugin-skill** GitHub 100+ stars
  Personal open-source project
  - Cut-centric agent film skill: shot transitions at the core, Blender blocking for space, then sound and editing.
  - Eight transition types with real-film examples; each shot's inputs viewable on the web and in ComfyUI.

Awards

- Chen Tianzhou Education Foundation Outstanding Student Scholarship 2025–2026
- Zhejiang Provincial Government Scholarship 2024–2025
- Zhejiang University Third-Class Scholarship 2023–2024

Skills

- **Frameworks**: SGLang, PyTorch, diffusers **Languages**: Python, C/C++ **Tools**: Linux/Shell, Docker, Git