Yongliang Wu

+86 15558561706 | yongliang0223@gamil.com | Homepage | Github | Google Scholar

Education

Southeast University, China. MS in Computer Science Sept 2023 – Apr 2026

- GPA: 3.60/4.0. Average score: 87.2. Award: Southeast University Graduate President Scholarship (Top 0.01%).

Southeast University, China. BS in Artificial Intelligence Sept 2019 – Apr 2023

- GPA: 3.75/4.0. Average score: 89.5. Award: Southeast University President Scholarship (Top 1%).

Selected Publications

KRIS-Bench: Benchmarking Next-Level Intelligent Image Editing Models
Yongliang Wu, et al. Advances in Neural Information Processing Systems, 2025.

Number it: Temporal Grounding Videos like Flipping Manga
Yongliang Wu, et al. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025.

VEU-Bench: Towards Comprehensive Understanding of Video Editing
Bozheng Li, ***Yongliang Wu***, et al. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025.

RSVP: Reasoning Segmentation via Visual Prompting and Multi-modal Chain-of-Thought
Yi Lu, Jiawang Cao, ***Yongliang Wu***, et al. Annual Meeting of the Association for Computational Linguistics, 2025.

Unlearning Concepts in Diffusion Model via Concept Domain Correction and Concept Preserving Gradient
Yongliang Wu, et al. AAAI Conference on Artificial Intelligence, 2025.

Video Repurposing from User Generated Content: A Large-scale Dataset and Benchmark
Yongliang Wu, et al. AAAI Conference on Artificial Intelligence, 2025.

Exploring Diverse In-Context Configurations for Image Captioning
Xu Yang, ***Yongliang Wu***, et al. Advances in Neural Information Processing Systems, 2023.

Internship Experiences

Research Intern, StepFun Inc. – Shanghai, China Feb 2025 – May 2025

- Responsible for constructing human motion image editing datasets to train the state-of-the-art Step1X-Edit image editing model. Curated synthetic datasets by extracting image pairs through sampling human motion variations from videos and leveraging video generation models, and produced corresponding editing instructions with a Vision-Language Model.

Research Intern, WeChat Vision, Tencent Inc. – Beijing, China June 2024 – Dec 2024

- Responsible for the training of a Multi-Modal Language Model WeMM-2B. Took charge of organizing pre-training data and use re-caption method to generate higher-quality image-text pairs. During the SFT stage, constructed a pipeline for automatic video annotation.

Machine Learning Engineer, OpusClip Inc. – Shanghai, China Nov 2023 – May 2024

- Responsible for developing an automatic short-form video generation system. Given a long-form video, the system can automatically select the highlight segments within it. Multiple short-form videos are then generated based on different topics at any ratio, and customized ones can also be produced according to the query provided by the user. The system is mainly LLM-driven and integrates several visual/audio foundation models as tools.

Machine Learning Engineer, Mettler-Toledo Inc. – Changzhou, Jiangsu, China Jul 2022 – Dec 2022

- Responsible for Training a lightweight classifier network on supermarket product images using MobileNet v3. Initialized with contrastive learning for pre-training, and trained with triplet and center losses in metric learning. Achieved high zero-shot classification accuracy on the out-of-domain test set.

Research Projects

Rectifying SFT for Better Generalization via an Unified SFT-RL View [Paper] [Code]

- We provide a theoretical analysis showing that SFT can be viewed as a policy gradient method with an ill-posed implicit reward, which explains its weak generalization ability. To address this, we propose Dynamic Fine-Tuning (DFT), a simple one-line modification that rescales the loss with token probabilities, effectively stabilizing updates and improving generalization. Experiments on multiple mathematical reasoning benchmarks and large models demonstrate that DFT significantly outperforms SFT and even rivals state-of-the-art offline and online RL methods. The paper is under submission.

Constructing Knowledge-based Reasoning Image Editing Benchmark [Paper] [Code]

- We introduce KRIS-Bench, a diagnostic benchmark designed to evaluate knowledge-based reasoning ability in instruction-guided image editing task. Grounded in educational theory, We categorize image editing tasks into Factual, Conceptual, and Procedural Knowledge types, and covers 22 representative tasks across 7 reasoning dimensions. Empirical results on 10 state-of-the-art iamge editing models reveal significant gaps in reasoning performance. The paper is accepted by NeurIPS 2025.

Enhance the Temporal Grounding Ability of Video-LLMs via Number Prompt [Paper] [Code]

- We propose a simple Visual Prompt method, which involves annotating specific numbers on each video frame to represent the frame number. It has been found that this method can significantly enhance the model's temporal grounding ability in a training-free manner, achieving SOTA results. The paper is accepted by CVPR 2025.

Unlearning Sensitive Concepts for Text-to-Image Diffusion Model [Paper] [Code]

- We propose a GAN-like adversarial training framework for unlearning the target concept in text-to-image diffusion model. Additionally, we introduce a gradient surgery approach, which serves to eliminate the conflict between the unlearn and retrain objectives. Through this, efficient unlearning is achieved while the utility of the model is maintained. The paper is accepted by AAAI 2025.

Exploring Multi-Modal In-Context Learning for Image Caption [Paper] [Code]

- We explore the problem of in-context examples selection in the Multi-Modal Large Language Model. For the visual and textual modalities respectively, we propose different selection strategies, achieving an average 20.9 CIDEr score improvement in the image caption task. The paper is accepted by NeurIPS 2023.

Competitions

Multilingual Video Reasoning challenge of the Vid-LLMs Workshop @CVPR2025.	*Winner*
Hour-long videoQA challenge of the Second Perception Test Challenge @ECCV2024.	*Winner*
Long-Term videoQA challenge of the LOVEU Workshop @CVPR2024.	*Winner*
Hour-long videoQA challenge of the Second Perception Test Challenge @ICCV2025.	*Runner-up*
Complex Video Reasoning challenge of the Vid-LLMs Workshop @CVPR2025.	*Runner-up*
Multi-View Foul Recognition challenge of the SoccerNet Challenge @CVPR2025.	*Third Place*

Academic Services

Conference Reviewer: ICCV 2025; CVPR 2025; ACMMM 2024-2025.

Journal Reviewer: IEEE TCSVT; FCS.

Workshop Organizer: NextVid Workshop@NeurIPS 2025.