Open-World Multimodal Understanding and Generation with Efficiently Finetuned Foundation Models

Long Chen

The Hong Kong University of Science and Technology
longchen@ust.hk

Background. With the astonishing abilities of different pretrained foundation models (*e.g.*, large language models (LLMs), vision-language models, diffusion models (DMs)), today's AI research tendency has been revolutionized. In this talk, I will answer two main questions: **Q1**: How can we efficiently train or fine-tune foundation models? **Q2**: How can we build strong open-world multimodal understanding and generation models with these pretrained foundation models?

Efficiently Finetune Foundation Models (Q1)

Parameter-Efficient Tuning (Chen 2024a). We argue that existing LoRA merging methods have "unrealistic" assumptions about the input data distribution. Thus, we propose a novel optimization-based method for LoRA merging.

Memory-Efficient Tuning (Diao 2024b,a). We argue that the scalability, adaptability, and generalizability of SOTA parameter-efficient transfer learning (PETL) methods are hindered by structural dependency and pertinency on specific pretrained backbones. Thus, we propose some memory-efficient PETL strategies to mitigate these weaknesses.

Modality-Efficient Tuning (Yu 2024). Existing fine-tuning methods rely heavily on extensive modal-specific pretraining and joint-modal tuning, leading to significant computational burdens for new modalities. In this work, we propose a flexible and scalable framework that enables MLLMs to continually evolve on modalities for X-modal reasoning.

Annotation-Efficient RLHF (Li 2024a). RLHF harnesses feedback, sourced either from humans or AI, as direct rewards or to shape reward models that steer language model (LM) optimization. We explore a methodology for reward composition that enables simultaneous improvements in LMs across multiple dimensions.

Multimodal Understanding and Generation (Q2)

Open-World Perception: We worked on two directions: 1) Using DM for data augmentation (Wang 2024a). In this paper, we analyze today's diffusion-based DA methods, and argue that they cannot take account of both faithfulness and diversity, which are two critical keys for generating high-quality samples and boosting classification performance. 2) Inspired by our humans, we first use LLMs to generate detailed and informative descriptions for different components of relation categories, such as subject, object, and spatial. These descriptions are used as description-based prompts for CLIP, enabling it to focus on specific visual features.

Multimodal Reasoning (Lin 2023; You 2023) Inspired by the success of LLMs on complex reasoning, we first conduct abductive reasoning with LLMs. Then we propose a novel generative framework to learn reasonable thoughts from LLMs for better multimodal reasoning.

Visual Generation and Editing. With the powerful pretrained DM, we also design a series of visual generation works, including image editing (Wang 2024b), video generation (Li 2025), and 3DGS editing (Wang 2025)

Multimodal Generation. To purse interleaved multimodal generation, we also collected a new benchmark (Chen 2024b). Meanwhile, we wrote a comprehensive survey to cover all multimodal benchmarks (Li 2024b).

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

References

Chen, H. 2024a. IterIS: Iterative Inference-Solving Alignment for LoRA Merging. *arXiv*.

Chen, W. 2024b. CoMM: A Coherent Interleaved Image-Text Dataset for Multimodal Understanding and Generation. *arXiv*.

Diao, H. 2024a. SHERL: Synthesizing High Accuracy and Efficient Memory for Resource-Limited Transfer Learning. In *ECCV*.

Diao, H. 2024b. Unipt: Universal parallel tuning for transfer learning with efficient parameter and memory. In *CVPR*.

Li, H. 2025. DisPose: Disentangling Pose Guidance for Controllable Human Image Animation. In *ICLR*.

Li, J. 2024a. Optimizing Language Models with Fair and Stable Reward Composition in Reinforcement Learning. In *EMNLP*.

Li, L. 2024b. A survey on multimodal benchmarks: In the era of large ai models. *arXiv*.

Lin, H. 2023. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. In *EMNLP*.

Wang, Y. 2024a. Inversion Circle Interpolation: Diffusion-based Image Augmentation for Data-scarce Classification. *arXiv*.

Wang, Y. 2025. View-consistent 3d editing with gaussian splatting. In *ECCV*.

Wang, Z. 2024b. Event-Customized Image Generation. *arXiv*.

You, H. 2023. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. In *EMNLP*.

Yu, J. 2024. LLMs can Evolve Continually on Modality for X-Modal Reasoning. In *NeurIPS*.