Towards Comprehensive Reasoning in Vision-Language Models

Yujun Cai[†] Jun Liu[§] Yiwei Wang[‡] Ming-Hsuan Yang[‡≀]
[†]University of Queensland [§] Lancaster University
[‡] University of California, Merced [≀] Google DeepMind
yujun.cai@uq.edu.au

1. Preference for Half-day Event and Justification for Four Speakers

We prefer a half-day event. The four proposers achieve high diversity in terms of gender, affiliation (four organizations), geography (three continents), seniority (two assistant professors, two full professors), and perspective (industry and academia). Also, their diverse expertise ensures broad coverage of the tutorial's related knowledge.

2. Course Description and Schedule

This tutorial will provide a comprehensive overview of reasoning capabilities in Vision-Language Models (VLMs), focusing on the transition from basic perception to complex inference. We will explore how multimodal reasoning differs from text-only reasoning, challenges specific to visual reasoning, and recent advances inspired by reasoning-focused LLMs. The tutorial aims to bridge the gap between theoretical foundations and practical implementations, offering attendees insights into both current capabilities and future research directions.

2.1. Introduction and Background

Recent large language models (LLMs) such as OpenAI's **O1** and **DeepSeek-R1** (R1) demonstrate that strong reasoning ability can be attained with innovative training [3]. These advanced *Reasoning LLMs* exemplify the new generation of language models capable of multi-step logical inference and self-correction. What sets these models apart is their ability to decompose complex problems into manageable steps and engage in sophisticated problem-solving strategies, often surpassing previous models on reasoning-intensive benchmarks.

Vision-Language Models (VLMs) have achieved remarkable progress in image captioning and visual question answering, yet genuine **reasoning** capabilities remain an open challenge. Unlike text-only LLMs, which can employ chain-of-thought reasoning, many VLMs still rely on pattern recognition and struggle with compositional logic. State-of-the-art VLMs often fail at seemingly simple tasks like accurately counting objects in an image when faced with clutter or occlusion [5]. This disconnect between impressive performance on standard benchmarks and persistent failures in reasoning-heavy scenarios highlights a fundamental limitation in current approaches.

In the vision-language domain, integrating such reasoning prowess is a critical frontier. This tutorial will survey the current state of reasoning in VLMs, highlighting key challenges and recent breakthroughs that move towards more *comprehensive reasoning*. We will particularly emphasize the unique difficulties posed by the visual modality, where information is presented spatially rather than sequentially, requiring models to develop specialized reasoning strategies that differ from those used in text-only contexts.

2.2. Enhancing Reasoning Capabilities in VLMs

A variety of methodologies have been proposed to bolster the reasoning capabilities of vision-language models. A prominent approach is to incorporate **Chain-of-Thought (CoT) reasoning** into multimodal tasks. Instead of directly predicting an answer, models generate intermediate **rationales** or step-by-step descriptions of their thought process. For example, the Multimodal-CoT framework separates vision-and-language reasoning into two stages (rationale generation and answer inference) [10]. Similarly, the **Visual CoT** dynamically focuses on key image regions and produces human-interpretable "thoughts" before answering [8]. These approaches not only improve performance on reasoning tasks but also enhance model transparency, allowing users to understand how the model arrived at its conclusions.

Another promising methodology is integrating external *reasoning engines* or using LLMs to guide VLM reasoning. One example is **Visual Program Distillation (VPD)**, which distills the problem-solving abilities of an LLM into a VLM [4]. This technique leverages the strength of specialized language models to guide the visual reasoning process, creating a synergistic relationship between the two modalities. Beyond VPD, other strategies include *reinforcement learning*, *self-consistency checks*, and *interactive reason-*

ing, where a model can ask follow-up questions before answering [1]. These approaches recognize that reasoning is often an iterative process that benefits from verification and refinement. By incorporating feedback mechanisms and allowing models to revise their initial hypotheses, these methods more closely mimic human reasoning patterns and produce more reliable outputs for complex visual reasoning.

2.3. Dataset Design for Vision-Language Reasoning

Robust datasets and benchmarks are pivotal for improving reasoning in VLMs. A key trend is the creation of benchmarks that demand *multi-step reasoning and world knowledge* beyond surface-level image recognition. For example, **Visual Commonsense Reasoning (VCR)** requires both an answer and a rationale, while **OK-VQA** forces models to recall external knowledge [2]. These benchmarks shift evaluation from simple pattern matching to more sophisticated reasoning processes that integrate visual perception with broader knowledge and inferential capabilities.

More recently, *evaluation suites* like **MMBench** and **MMMU** aggregate tasks across numerous domains [5]. MMMU consists of multimodal questions from college-level exams, requiring vision, expert-level knowledge, and reasoning. MMBench evaluates abilities from basic perception to high-level reasoning. Increasingly, researchers are moving beyond manually annotated datasets to develop methods for automatically generating high-quality reasoning examples at scale, potentially addressing the data bottleneck that has limited progress in this area.

2.4. Advancements in Model Architectures

Achieving stronger reasoning in VLMs requires **architectural innovations** that better fuse visual and textual information. One significant development has been the rise of *multimodal large language models (MLLMs)*. **BLIP-2** employs a lightweight querying transformer to bridge image features and a frozen language model. Approaches like **LLaVA [7], SPHINX [6], and Qwen-VL [9]** follow a two-stage strategy: encode the image, then feed embeddings into a language model. This modular approach allows each component to specialize in its respective domain while facilitating effective information exchange between modalities.

Further efforts focus on **object-level grounding** and *spatial feature mixing*. **Kosmos-2** allows the model to input bounding box descriptions, improving spatial reasoning. **Mixture of Features (MoF)** combines different feature types, enhancing fine-grained distinctions. These approaches recognize that reasoning often requires attending to specific objects and their relationships within an image, rather than processing the entire visual input as an undifferentiated whole. The ability to ground language in specific visual regions and vice versa represents a significant step toward more comprehensive reasoning capabilities.

2.5. Future Directions and Open Challenges

Despite rapid progress, truly **comprehensive reasoning** in vision-language models remains a frontier. One major challenge is *compositional generalization*—models often fail on novel combinations of concepts. This limitation suggests that current VLMs may be learning superficial correlations rather than developing a deeper understanding of underlying principles that would enable generalization to unseen scenarios. Another issue is *explainability and trustworthiness*. Models still produce incorrect justifications or confidently state wrong answers. This disconnect between confidence and correctness undermines trust in model outputs and limits their usefulness in critical applications. Developing VLMs that can accurately assess their own uncertainty and provide reliable explanations for their reasoning processes remains an important research direction.

Finally, we stress the importance of rigorous evaluations. Ultimately, the field aims to move from perception to *cognition*, creating systems that see, understand, reason, and explain. This transition requires not only technical advances but also a deeper understanding of how to measure and evaluate higher-order cognitive abilities in artificial systems. Developing more sophisticated evaluation frameworks that can distinguish between genuine reasoning and sophisticated pattern matching will be essential for guiding future research in this area.

2.6. Schedule and Topics

Half-Day Session (9:00-12:30):

- **9:00-9:30 — Introduction to Vision-Language Reasoning**
  Definitions, taxonomy of reasoning types, differences between perception and reasoning in VLMs, connections to reasoning-focused LLMs like O1 and DeepSeek-R1
- **9:30-10:15 — Enhancing Reasoning Capabilities in VLMs**
  Chain-of-thought techniques, Visual CoT, reasoning engines, Visual Program Distillation (VPD), interactive reasoning approaches
- **10:15-10:45 — Coffee Break & Demonstrations**
- **10:45-11:30 — Dataset Design & Model Architectures**
  Reasoning-focused benchmarks (MMBench, MMMU), multimodal architectures (Flamingo, BLIP-2, LLaVA), object-level grounding and spatial feature mixing
- **11:30-12:15 — Future Directions and Open Challenges**
  Compositional generalization, explainability, multimodal reasoning across video/3D/audio, integration of external knowledge, model efficiency
- **12:15-12:30 — Q&A and Closing Remarks**

3. Subject Areas

Primary: Reasoning Vision-Language Models
Secondary Subject Areas:
- Multimodal Large Language Models
- Chain-of-Thought and Explicit Reasoning
- Benchmarking and Evaluation of Reasoning Capabilities

4. Expected Target Audience

Composition: Researchers, practitioners, and students interested in multimodal deep learning, particularly focusing on reasoning capabilities beyond pattern recognition. The tutorial is designed to benefit both those with experience in vision-language modeling seeking to enhance reasoning aspects, and those from adjacent fields (e.g., NLP or cognitive science) looking to understand the unique challenges of visual reasoning.
Estimated Attendance: Large (>300 attendees)

5. Format

This tutorial will use a mixed in-person format. The primary instructors will deliver content in-person, while supporting virtual attendance through Live streaming of all sessions with moderated Q&A.

6. Relation to Recent Tutorials

This tutorial builds upon yet significantly differs from recent related tutorials:
- CVPR 2024: "Recent Advances in Vision Foundation Models" focused broadly on vision foundation models but dedicated limited time to reasoning capabilities.
- ECCV 2024: "Large Multimodal Foundation Models" covered general MLLM architectures but not the specialized reasoning mechanisms we will explore.

Our tutorial uniquely focuses on the reasoning frontier in VLMs, with special emphasis on techniques to enhance logical inference, spatial understanding, and multi-step problem solving. Unlike previous tutorials, we explicitly connect advances in LLM reasoning (e.g., those found in models like DeepSeek-R1) to the visual domain, providing practical guidance on implementing and evaluating reasoning capabilities.

7. Links to Previous Recorded Talks

Ming-Hsuan Yang's talk at Microsoft Research
Ming-Hsuan Yang's talk at UCF
Yiwei Wang's talk at University of Queensland

8. Materials and Resources

Attendees will receive comprehensive slide decks covering all tutorial segments and a gitHub repository with code implementations of key reasoning techniques. All materials will be hosted on a dedicated website that will remain available for at least one year after the tutorial.

9. Special Requirements

No special requirements.

Relevant Publications

[1] Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Hao Zhang, and Chuang Gan. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. *arXiv preprint arXiv:2301.05226*, 2023. 2

[2] Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1254–1262, 2024. 2

[3] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1

[4] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9590–9601, 2024. 1

[5] Hang Hua, Yunlong Tang, Ziyun Zeng, Liangliang Cao, Zhengyuan Yang, Hangfeng He, Chenliang Xu, and Jiebo Luo. Mmcomposition: Revisiting the compositionality of pre-trained vision-language models. *arXiv preprint arXiv:2410.09733*, 2024. 1, 2

[6] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 2

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

[8] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv e-prints*, pages arXiv–2403, 2024. 1

[9] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[10] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 1

10. Proposers' Biographies

Yujun Cai is a Lecturer in the University of Queensland in Australia. Before that, she was a Research Scientist in Meta Reality Lab. Her research lies in multi-modal human perception, vision-language models, and natural language processing. She obtained her PhD. degree from Nanyang Technological University in Singapore. Additional information is available at https://vanoracai.github.io/.

Jun Liu is a Professor in Lancaster University, UK. His recent research focuses on multi-modal data analysis, video understanding and generation, and privacy and safety of large models. He obtained the best paper awards from PREMIA in 2016 and 2019, the Best Doctoral Thesis Award from EEE at NTU in 2020, the IEEE VSPC Rising Star Honorable Mention Award in 2024. He is a Senior Area Editor of IEEE Transactions on Image Processing. He has been a co-organizer of around 10 international workshops. Additional information is available at https://wp.lancs.ac.uk/vl/.

Yiwei Wang is an Assistant Professor in the Computer Science Department of University of California, Merced. His research lies in natural language processing and graph machine learning. His recent research focuses on the controllable generation of LLMs and applications of LLMs in healthcare. He was awarded the 2021 SDSC Research Fellowship. Additional information is available at https://wangywust.github.io/.

Ming-Hsuan Yang is a Professor of Electrical Engineering and Computer Science at UC Merced, with a Ph.D. from UIUC (2000). His research spans computer vision, pattern recognition, AI, robotics, and machine learning. He has served as Program Chair for ICCV and ACCV, and as Associate Editor-in-Chief of IEEE PAMI. Yang has been an Area Chair for major conferences including CVPR, ICCV, ECCV, NeurIPS, ICLR, and ICML. His notable recognitions include the ICML 2024 Best Paper, Longuet-Higgins Prize at CVPR 2023, NSF CAREER Award (2012), and status as a Highly Cited Researcher (2018-2024). He is a Fellow of IEEE, ACM, and AAAI.