

Toward Trustworthy Vision-language Models in the Wild: Theory, Algorithm and Application

Bo Peng*
Computer Science
University of Technology Sydney
Sydney, Australia
Bo.Peng-7@student.uts.edu.au

Xuefeng Du
Nanyang Technological University
Singapore, Singapore
xuefeng.du@ntu.edu.sg

Zhen Fang
University of Technology Sydney
Sydney, Australia
Zhen.Fang@uts.edu.au

Abstract

Vision-Language Models (VLMs) have revolutionized multimedia applications by enabling open-vocabulary search and generation. While significant strides have been made in developing powerful VLMs, the field remains in the early stages of rigorously evaluating and understanding their real-world vulnerabilities, both empirically and theoretically. In response to this important but under-explored gap, the workshop aims to provide a forum for researchers to share advanced in theories, algorithms and applications of trustworthy multi-modal learning, fostering diverse viewpoints on core principles and emerging techniques for developing trustworthy VLMs in the wild. The workshop provides a platform for researchers to showcase their work, exchange ideas, and foster potential collaborations. Additionally, it serves as a valuable opportunity for practitioners to stay abreast of the latest developments in developing trustworthy vision-language models.

CCS Concepts

• **Computing methodologies** → **Machine learning**; **Computer vision**; **Natural language processing**.

Keywords

Vision-language Models, Trustworthy Machine Learning

ACM Reference Format:

Bo Peng, Xuefeng Du, and Zhen Fang. 2026. Toward Trustworthy Vision-language Models in the Wild: Theory, Algorithm and Application. In *International Conference on Multimedia Retrieval (ICMR '26)*, June 16–19, 2026, Amsterdam, Netherlands. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3805622.3811235>

1 Introduction

Vision-Language Models (VLMs), such as CLIP [5], LLaVA [4], and GPT-4V [11], have revolutionized multimedia applications, such as captioning [10], visual question answering [3], cross-modal retrieval [2], image clustering [6, 8, 9], and OOD learning [1, 7], by enabling open-vocabulary search and generation. At the same time, the community is encountering a growing gap between benchmark

performance and deployment reliability in real-world, open-domain settings—what this workshop calls “in the wild.”

This workshop treats **trustworthiness** as a systems-level requirement for VLM-enabled multimedia applications. Specifically, we frame trustworthiness as a multi-dimension property spanning:

- (1) **Reliability**: Stability under distribution shift; resilience to noisy, incomplete, or imperfect multi-modal inputs.
- (2) **Truthfulness**: Reducing hallucinations by grounding model outputs in visual evidence and/or retrieved provenance.
- (3) **Calibration**: Knowing when to abstain, defer, or request additional information—especially under ambiguity, uncertainty, or insufficient evidence.
- (4) **Safety**: Adversarial robustness, prompt-injection resistance in multimodal agents.
- (5) **Responsible deployment**: Bias and fairness, explainability, privacy, legal and ethical constraints.

While significant strides have been made in developing powerful VLMs, the field remains in the early stages of rigorously evaluating and understanding their real-world vulnerabilities, both empirically and theoretically. In response to this important but under-explored gap, the workshop aims to provide a forum for researchers to share advanced in theories, algorithms and applications of trustworthy multi-modal learning, fostering diverse viewpoints on core principles and emerging techniques for trustworthy VLMs in the wild.

2 Topics of Interest

The workshop will encompass a broad spectrum of topics including but are not limited to:

A. Theory:

- Uncertainty quantification for VLMs.
- Theoretical bounds on VLMs under distribution shifts.
- Causal perspectives on spurious correlations and multi-modal shortcut learning.
- Robust optimization and generalization bounds for multi-modal learning objectives

B. Algorithms:

- Out-of-distribution (OOD) generalization, OOD detection and test-time adaptation
- Detecting and mitigating hallucination from VLMs
- Robust training and inference against adversarial attacks, prompt injection and imperfect inputs.
- Post-hoc methods for explaining VLM decisions.
- Techniques for auditing and debiasing vision-language datasets.

*Correspondence to Bo Peng (Bo.Peng-7@student.uts.edu.au)



This work is licensed under a Creative Commons Attribution 4.0 International License. *ICMR '26, Amsterdam, Netherlands*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2617-0/26/06
<https://doi.org/10.1145/3805622.3811235>

Table 1: Tentative schedule of this workshop.

Activity	Duration	Speaker	Organization
An opening of the workshop	5 min	Bo Peng	University of Technology Sydney
Invited Talk 1	30 min	Yingjun Du	University of Amsterdam
Invited Talk 2	30 min	Mengyue Yang	University of Bristol
Invited Talk 3	30 min	Pascal Mettes	University of Amsterdam
Coffee Break	10 min		
Paper1 Presentation	15 min	TBD	
Paper2 Presentation	15 min	TBD	
Paper3 Presentation	15 min	TBD	
Paper4 Presentation	15 min	TBD	
Paper5 Presentation	15 min	TBD	

- Methods to ensure fairness in the context of VLMs.
- Privacy-preserving machine learning for VLMs.

C. Benchmark and Application:

- Trustworthy usages of VLMs for scientific domains, including program languages, climate science, healthcare, life sciences, physics, and cognitive science.
- Detecting AI-generated content and deepfakes.
- New benchmarks and evaluation protocols for real-world trustworthiness

3 Organizers

Bo Peng is a final-year Ph.D. candidate at University of Technology Sydney in computer sciences. His works develop both empirical algorithms and theoretical guarantees for trustworthy machine learning, which led to several publications at major top-tier conferences, such as NeurIPS, ICML, ICLR, ICCV and KDD. He received the NeurIPS 2025 Outstanding Reviewer Award.

Xuefeng Du is currently an Assistant Professor at College of Computing and Data Science, Nanyang Technological University, Singapore. His research focuses on responsible machine learning and foundation models. His works develop both empirical algorithms and theoretical understandings to enable open-world learning that can function safely and adaptively in the presence of evolving and unpredictable data streams, which led to several publications at major machine learning conferences, such as NeurIPS, ICML, and ICLR. He has received the Jane Street Graduate Research Fellowship and Rising Stars in Data Science award.

Zhen Fang is an Australian DECRA Fellow. He is a lecturer at the University of Technology Sydney. His research interests center on the theory of machine learning models, aiming to enhance the responsibility of machine learning models in the wild. He has published several papers related to responsible machine learning in top-tier conferences and journals, e.g., ICML, NeurIPS, ICLR, IEEE TNNLS, IEEE TCYB, IEEE TKDE, JMLR and IEEE TPAMI. He received the NeurIPS 2022 Outstanding Paper Award.

4 Tentative Schedule

We plan to hold a hybrid format of workshop, i.e., both onsite and online. For the onsite one at least two organizers will attend in-person to host the workshop. The workshop will include two major activities, the invited keynotes, and the paper presentations. We will invite keynote presentations for a half-day workshop, following by 5 accepted workshop presentations. The speakers are the experts on the relevant community from different organizations globally. The schedule of the workshop are listed in Table 1

References

- [1] Peng Bo, Lu Jie, Guangquan Zhang, and Fang Zhen. 2026. Debaised Negative Mining Improves Out-of-distribution Detection with Pre-trained Vision-Language Models. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*. 1171–1181.
- [2] T P Kalyan, Piyush S Pasi, Sahil N Dharod, Azeem A Motiwala, Preethi Jyothi, Aditi Chaudhary, and Krishna Srinivasan. 2024. Wikido: A new benchmark evaluating cross-modal retrieval for vision-language models. *Advances in Neural Information Processing Systems* 37 (2024), 140812–140827.
- [3] Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. 2025. Natural language understanding and inference with mllm in visual question answering: A survey. *Comput. Surveys* 57, 8 (2025), 1–36.
- [4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.
- [5] Peyman Morteza and Yixuan Li. 2022. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 7831–7840.
- [6] Bo Peng, Yuanwei Hu, Bo Liu, Ling Chen, Jie Lu, and Zhen Fang. 2026. Delving into Spectral Clustering with Vision-Language Representations. *arXiv preprint arXiv:2602.09586* (2026).
- [7] Bo Peng, Jie Lu, Guangquan Zhang, and Zhen Fang. 2025. An Information-theoretical Framework for Understanding Out-of-distribution Detection with Pretrained Vision-Language Models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- [8] Bo Peng, Jie Lu, Guangquan Zhang, and Zhen Fang. 2025. On the Provable Importance of Gradients for Autonomous Language-Assisted Image Clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19805–19815.
- [9] Bo Peng, Jie Lu, Guangquan Zhang, and Zhen Fang. 2025. On the Provable Importance of Gradients for Language-Assisted Image Clustering. *arXiv preprint arXiv:2510.16335* (2025).
- [10] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. 2023. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems* 36 (2023), 40924–40943.
- [11] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421* (2023).