氏名	Yosuke Miyanishi		
学位の種類	博士（情報科学）		
学位記番号	博情第 561 号		
学位授与年月日	令和 7 年 9 月 24 日		
論文題目	INTERPRETING AND MITIGATING PRESENTATION BIASES IN VISION-AND-LANGUAGE TRANSFORMERS		
論文審査委員	Nguyen Le Minh	JAIST	Professor
	Shogo Okada	JAIST	Professor
	Kiyoaki Shirai	JAIST	Professor
	Naoya Inoue	JAIST	Associate Professor
	Tomoko Matsui	ISM	Professor

論文の内容の要旨

Upon the remarkable success achieved by Vision-and-language models (VLMs) in various tasks such as Visual Question Answering (VQA) and Hateful Meme Detection (HMD), the alignment of VLMs with human values and semantic intent is essential for their safe and reliable deployment across complex multimodal tasks. However, they remain vulnerable to presentation biases: systematic errors arising from how multimodal inputs are presented rather than their semantic content. This dissertation develops a unified, interpretability-driven framework to detect, quantify, and mitigate two central presentation biases:

1. Modality Bias: We introduce two novel causal metrics—Multimodal Intersectional Treatment Effect (MITE), which measures bias in predicted hatefulness, and Modality Interaction Disentangled Attribution Score (MIDAS), which assesses bias in attention attributions—to pinpoint overreliance on text or image cues within pretrained BERTbased HMD models. Leveraging these insights, we design a posthoc calibration pipeline that re-ranks predictions based on attention attributions, improving detection robustness (up to +2.7 points in accuracy for the Hateful Memes Challenge (HMC) dataset).

2. Formatting Bias: We propose Representational Shift Theory (RST), a theoretical account of how In-Context Learning (ICL) affects latent space under non-semantic input variations (e.g. the number of images/conversation turns). Guided by RST, we validate two ICL interventions: CLIP-based example selection for VQA and counterfactual prompting for HMD. Empirical evaluations on six standard VQA benchmarks and the HMC dataset demonstrate significant gains in semantic understanding and format-robust performance, especially in challenging VQA datasets (four out of six datasets, up to +15 points

in accuracy) and text-dominant tasks (≃ +0.8 points in F1 score for HMD).

Together, these contributions advance the interpretability toolkit for VLM alignment by providing actionable metrics and interventions to build more reliable, unbiased multimodal AI systems. I conclude by discussing limitations in scalability and generalizability, and outline future directions toward mechanistic interventions and real-world deployment of aligned VLMs.

Keywords: Gradient, VLM, Transformer, Bias, In-Context Learning.

論文審査の結果の要旨

This dissertation addresses the challenge of aligning vision–language models (VLMs) with human intent while guarding against presentation biases—errors driven by how inputs are formatted rather than what they mean.

It presents a cohesive, practical, and unified, interpretability-driven framework that:

(i) detects and quantifies modality bias using two causal, actionable metrics—MITE (outcome bias) and MIDAS (attribution bias);

(ii) and (ii) mitigates formatting bias via the proposed Representational Shift Theory (RST) and two simple in-context learning interventions (CLIP-guided example selection for VQA and counterfactual prompting for HMD).

(iii) Experiments across multiple VQA benchmarks and the Hateful Memes Challenge show clear, reproducible gains in robustness and accuracy.

Together, these contributions expand the interpretability toolkit for VLM alignment by providing actionable metrics and practical mitigation strategies for building more reliable, less biased multimodal systems. The dissertation also discusses limitations in scalability and generalizability and outlines future work on mechanistic interventions and real-world applications**.**

As a result of this work, the candidate has published in a leading conference in the field of Natural Language Processing**.** The candidate also open-source code and data for replicating his experimental results.

In summarily, this is a strong, well-argued, and experimentally solid thesis that moves the field toward **more reliable and unbiased multimodal AI**. It provides clear concepts, useful tools, and measurable improvements. Overall, this is an excellent dissertation, and we approve of awarding a doctoral degree to Mr. Yosuke Miyanishi.