

Title	視覚言語トランスフォーマにおける表現バイアスの解釈と軽減
Author(s)	宮西, 洋輔
Citation	
Issue Date	2025-09
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/20080
Rights	
Description	Supervisor: NGUYEN, Minh Le, 先端科学技術研究科, 博士

Doctoral Dissertation

INTERPRETING AND MITIGATING PRESENTATION BIASES IN
VISION-AND-LANGUAGE TRANSFORMERS

YOSUKE MIYANISHI

Supervisor NGUYEN LE MINH

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

September 2025

Abstract

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 BERT-based HMD models. Leveraging these insights, we design a post-hoc 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 (\simeq +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.

Acknowledgment

I thank my supervisors in JAIST for their insightful revisions for this thesis. I especially thank Prof. Nguyen Minh Le for his continued support throughout my student period. I thank Dr. Wensheng Lin for open-sourcing the useful template¹ used to write this thesis. I thank my CyberAgent² colleagues for their understanding of the significance of Ph.D. course completion, and their financial support allowing me to attend the international conferences. I thank Dr. Takehito Yoshiki (Gree Inc.) for his insightful comments. I thank ALIGN network³ for the collaboration opportunities with the safety-concerned machine learning researchers and other domain experts. Whether my submissions were (will be) accepted or rejected, I thank anonymous academic conference reviewers for their valuable feedback. Last but not least, I thank my family; I thank my father Prof. Takayuki Miyanishi (Nagasaki Univ.) for being a role model for my interest-driven career, and my mother Mami for her continuous support in the early stage of my life. I thank my wife Risa and my newborn daughter Hirona for their support for my multitasking life as a data scientist, a research unit manager⁴, a Ph.D. student, and their husband/father.

¹<https://github.com/linwest/JAISTthesisLaTeX>

²<https://www.cyberagent.co.jp/>

³<https://www.aialign.net/>

⁴<https://www.cyberagent.co.jp/techinfo/info/detail/id=25611>

List of Abbreviations

M^2 OoD	Multi-image Multi-turn OoD
attr	Attention attribution score
BERT	Bidirectional Encoder Representations from Transformers
CFP	CounterFactual Prompting
CLIP	Contrastive Language-Image Pre-Training
CNN	Convolutional Neural Network
GPT	Generative Pre-trained Transformer
HM	Hateful Meme
HMC	Hateful Memes Challenge
HMD	Hateful Meme Detection
ICL	In-Context Learning
ID	In-Distribution
LLaVA	Large Language and Vision Assistant
LLM	Large Language Model
LM	Language Model
MI	Mechanistic Interpretability
MIDAS	Modality Interaction Disentangled Attribution Score
MiMiC	Minimally Modified Counterfactuals
MITE	Multimodal Intersectional Treatment Effect
NLP	Natural Language Processing
OoD	Out-of-Distribution
RepE	Representation Engineering
RNN	Recurrent Neural Network
RST	Representational Shift Theory
SimVLM	Simple Visual Language Model
SM OoD	Single-image Multi-turn OoD
SotA	State-of-the-Art
TE	Treatment Effect
VL	Vision-and-Language
VLM	Vision-and-Language Model
VQA	Visual Question Answering
XAI	EXplainable Artificial Intelligence
ZSL	Zero-Shot Learning

List of Figures

1.1	Schematic overview of the entire dissertation. (A,B) Presentation biases studied in this dissertation. (A) Modality bias in Hateful Meme Detection (HMD; Chapter 3). A State-of-the-Art (SotA) HMD model Vilio [1] suffers from the bias towards text modality, potentially due to the shortage of multimodal training. (B) Formatting bias in Visual Question Answering (VQA) (Chapter 4). Two variants of LLaVA [2], which exhibit SotA performance in VQA and are trained solely with single-image and single-turn vision-and-language resources, struggle with multi-image and/or multi-turn VQA queries. (C,D) High-level overview of the proposed methodologies. (C) Approach for modality bias in HMD. We propose two metrics for quantifying the modality biases in a model’s predictions and in attention attribution [3]. Given the insights about the biases within the model, we calibrate the prediction with attention attribution for debiasing the base predictor. (D) Approach for formatting bias in VQA. First, we propose Representation Shift Theory (RST), a novel theory about In-Context Learning (ICL) for formatting bias mitigation. Next, based on RST, we empirically validate two hypotheses for the bias mitigation: 1) prompt engineering can contribute to the performance gain in a text-predominant task (validated by counterfactual prompting for HMD), and 2) careful VQA example selection can lead to better performance when the model struggles with the task in the zero-shot setting (validated by CLIP-based example selection for VQA).	6
-----	---	---

3.1	Schematic overview of the proposed pipeline. First, a hateful meme and its benign confounder counterpart are given to Vilio for extracting its predicted hatefulness and the attention attribution (top). The confounders are divided into image benign (the image is replaced from the original hateful meme so that it flips the label to benign) and text benign (the text is replaced, flipping the label). In Experiments I and II (middle), we quantify modality bias as the prediction difference between a hateful example and its benign confounder counterpart (MITE; Experiment I) and as the attention attribution difference (MIDAS; Experiment II). To see if attention attribution is effective for calibrating the prediction upon modality bias or not, we then build a LightGBM re-ranker (bottom), solely with the prediction or with the prediction and the attention attribution.	23
3.2	MITE across Oscar (O: left), UNITER (U: middle), and VisualBERT (V: right), shown separately for <i>Image-to-Image</i> (magenta) and <i>Text-to-Text</i> (cyan) settings. Larger MITE values indicate greater reliance on joint text-image signals. A notable gap between the two bars suggests asymmetric reliance on one modality.	29
3.3	Global MIDAS for Text-to-Text (left) and Image-to-Image (right) samples, for Oscar (top), UNITER (second row), VisualBERT (third row), and VisualBERT (text-only pretrained) (bottom). Bars in each graph show attr summed over all the elements, $MIDAS_{\text{text}}$, $MIDAS_{\text{cross}}$, and $MIDAS_{\text{image}}$ (from left to right). Consistently with MITE result, VisualBERT shows the strongest reliance on $MIDAS_{\text{text}}$ for both tasks, while other models relies more on $MIDAS_{\text{image}}$ and $MIDAS_{\text{cross}}$ for the Image-to-Image task.	30
3.4	Local MIDAS explanations for a representative hateful (row 1), text-benign (row 2), and image-benign (row 3) sample using UNITER. Top image regions and text tokens receive the highest attribution under each intervention. UNITER attends to "dishwasher" in the text and a woman in the image, suggesting that this model successfully capture the hateful analogy.	31

4.1	Schematic Overview of Representational Shift Theory (RST) with <i>Banana Mustache</i> test input and <i>Tomato Nose</i> In-Context Learning (ICL) example. Both in a zero-shot setting (ZSL; black line) and in ICL (red line), an instruction (question) is first provided with LLaVA to compose the representation H_{inst} . Next, in a zero-shot setting, the test input forms the meta-gradient $\Delta W_{\text{inst/ZSL}}$, resulting in the test-input’s representation H_{ZSL} . Finally, the unembedding weight W_{emb} converts H_{ZSL} to the output Y_{ZSL} . In contrast, in ICL, an example is inserted between the instruction and the test input to contextualize the semantic components (analogy of fruits and facial parts) to form the meta-gradient $\Delta W_{\text{inst/icl}}$ and $\Delta W_{\text{icl/zsl}}$, and then the ICL example’s representation H_{ICL} and the output Y_{ICL} . We argue that the difference of test-input representation $H_{\text{ICL}} - H_{\text{ZSL}}$ (input shift) and that of output representation $Y_{\text{ICL}} - Y_{\text{ZSL}}$ (output shift) are the core of ICL. .	43
4.2	Schematic Drawings of Hypotheses I/II and Experiments I/II. (A) Hypothesis I and Experiment I. We hypothesize that semantically rich ICL example given in an unfamiliar format contributes to the performance when the zero-shot performance is poor (left). To validate this hypothesis, we tested LLaVA’s Out-of-Distribution (OoD) ICL performance by adding an ICL example most similar to the test input based on CLIP embedding (semantically rich) to the LLaVA variants <i>not</i> trained with multi-image datasets (right). When the model successfully mitigate formatting bias, it extracts the <i>semantic</i> analogy despite the unseen format (large and explicit <i>formatting bias</i>). (B) Hypothesis II and Experiment II. When LLaVA is textually biased towards the hateful label (top), we hypothesize that enhancing text-to-text interaction facilitates the unbiased decision (bottom), and test this hypothesis with counterfactual prompting (instructing the model to propose a caption to compare with the original caption).	44
4.3	The representative counterfactual prompt. The system prompt is truncated for illustrative purposes. The curly brackets indicate the placeholders. This prompt instructs the model to generate a neutral caption that fits with the given image, and then to use that caption as a contrastive example for the hateful meme detection.	48

4.4	Performance summary of LLaVA-Llama2. zsl and icl represent Zero-Shot Learning (ZSL) and Out-of-Distribution In-Context Learning (OoD ICL; denoted simply as <i>ICL</i> in the figure legends for brevity). OoD ICL results in better performance for four datasets, where the zero-shot performances are poor. .	51
4.5	A representative LLaVA-Llama2’s response from the GQA dataset. (A) LLaVA-Llama2 does not follow a part of the instruction where it is required to answer the question in a single word. (B) LLaVA-Llama2 responded in a single word with an ICL example.	58
4.6	A representative LLaVA-1.5’s response for which OoD ICL impacts the performance positively from the TextVQA dataset. (A) LLaVA-1.5 answered the monetary value question without its unit (<i>just 1, without penny</i>), potentially due to the missing OCR token for the unit. (B) LLaVA-1.5 answered correctly , potentially because the ICL example attributed one penny to the value of the copper coin.	59
4.7	A representative LLaVA-1.5’s response when the ICL example is randomly sampled from the training dataset. Seemingly irrelevant image-text ICL example does not affect the model’s response.	60
4.8	Representational shift across the learning type. Each entry is the similarity of the input between two conditions. For example, the left-top value 0.173 is the similarity of the input between hateful samples of a CFP setting and benign samples of a ZSL setting. While the hateful samples and the benign samples are similar for ZSL and ICL (i.e., M^2 OoD) settings, CFP hateful samples and benign samples are less similar. . . .	61
A.1	Sampled derived from Oscar.	69
A.2	Samples derived from VisualBERT.	70
A.3	Oscar $MIDAS_{att}$	71
A.4	UNITER $MIDAS_{att}$	71
A.5	VisualBERT $MIDAS_{att}$	72
A.6	Oscar $MIDAS_{grad}$	72
A.7	UNITER $MIDAS_{grad}$	73
A.8	VisualBERT $MIDAS_{grad}$	73
B.1	The performance summary of LLaVA-1.5. OoD ICL dropped the performance, suggesting the rich semantics in the test input. .	79

- B.2 A representative LLaVA-1.5’s response for which OoD ICL impacts the performance *negatively* from the TextVQA dataset. (Left) LLaVA-1.5 answered the **correct** object tag (**ale**). (Right) LLaVA-1.5 answered *incorrectly*, potentially caused by the wrong label (*pine* and *belt*) presented in the ICL example. 83
- B.3 Performance summary of InternVL. MM OoD dropped the performance for all the datasets, potentially reflecting that the baseline performance is moderate to high for all the datasets. . 84

List of Tables

1.1	Definitions of key terms.	1
3.1	Key terms used in Chapter 3.	15
3.2	Breakdown of all hateful and benign pairs. Each hateful meme is paired with one provided benign example and three picked confounders to ensure both modalities are covered.	26
3.3	Summary of the three models used in our experiments.	27
3.4	Calibration results. The first column lists the model, followed by the submodel identifier. The next three columns show AUROC scores for the raw prediction only, single attention attribution [3], and two score baseline [4]. The last column shows the performance of our proposed three-interaction-type method. Performance is evaluated by AUROC, shown in percentage for brevity. For Oscar and VisualBERT, our three-interaction-type method outperforms single or dual attention split baselines with an exception of O36.	32
4.1	Definitions of key terms.	37
4.2	Regression Coefficient*100 of the mixed effect model’s prediction with the dummy variables representing the datasets and the models. The prediction shows a much higher coefficient than the dummy variables, validating our models.	52
4.3	Accuracy of the on-the-fly context selection. Bold indicates the best performance in each row. In the LLaVA-Llama2 case, the performance is bounded by the OoD ICL (as with the figures, denoted as <i>ICL</i> in the table legends) accuracy since ZSL performance is extremely low. In the LLaVA-1.5 case, it outperforms <i>both ZSL and ICL</i> , suggesting its efficiency when the performance of the two strategies is comparable.	53
4.4	Hateful memes detection performance. CFP’s performance is better than ZSL while regular ICL (M^2 OoD) dropped the performance, supporting Hypothesis II.	54

B.1	Distribution of response lengths for correct vs. incorrect predictions made by LLaVA-Llama2 for the GQA dataset. Note the high occurrence of single-word responses in correct predictions.	79
B.2	Weight means and standard deviations ($\times 10\,000$) for random-effect only model vs. mixed effect model.	80
B.3	Regression coefficients of the variables representing model (LLaVA 1.5 or LLaVA-Llama2), dataset, and presence/absence of ICL examples. <i>all</i> represents the result of an all-variable model. R^2 values are multiplied by 100 for brevity. The result only with the model variable is similar to the all-variable model, reflecting the performance gap between LLaVA 1.5 and LLaVA-Llama2 (Fig. 4.4 and B.1).	81
B.4	Impact of multi-image ICL in GQA for InternVL 1b. N steps indicate the number of inference steps. The numbers with an error represent accuracy(%) in the corresponding setting. ICL boosted the performance when the number of steps was above six, implying that the ICL positively affects the performance when the task is challenging.	81

Contents

Abstract	I
Acknowledgment	III
List of Abbreviations	V
List of Figures	VII
List of Tables	XIII
Contents	XV
Chapter 1 Introduction	1
1.1 Terminology	1
1.2 Background and Motivation	2
1.3 Problem Statement	4
1.4 Research Objective	5
1.5 Dissertation Outline	5
Chapter 2 Related Work	7
2.1 Outline	7
2.2 LMs and LLMs	7
2.3 VQA	8
2.4 HMD	8
2.5 Bias Detection and Mitigation	9
2.5.1 Modality Bias	9
2.5.2 Formatting Bias	10
2.6 Interpretability	11
2.6.1 Causally Interpreting LLMs	11
2.6.2 Interpretability for Debiasing LLMs	11
2.6.3 VLM Applications	12
Chapter 3 Modality Bias within Hateful Meme Detectors	15

3.1	Terminology	15
3.2	Chapter Summary	15
3.3	Introduction	16
3.4	Related Work	17
3.4.1	Intersectionality and Causal Analyses	17
3.4.2	Explainability for Hateful Meme Models	17
3.4.3	Adversarial Robustness and Countermeasures	18
3.4.4	Datasets and Multimodal Architectures	18
3.4.5	Broader Bias and Debiasing Approaches	19
3.5	Methodology	19
3.5.1	Preliminaries	19
3.5.2	Proposed Methodology Overview	21
3.5.3	Multimodal Intersectional Treatment Effect (MITE)	22
3.5.4	Modality Interaction Disentangled Attribution Scores (MIDAS)	23
3.5.5	Bias-Aware Calibration	24
3.6	Experimental Settings	25
3.6.1	Dataset and Confounder Extraction	26
3.6.2	Models	27
3.6.3	Bias-Aware Calibration Set-up	28
3.6.4	Implementation Details	28
3.7	Results & Discussion	28
3.7.1	Experiment I: Modality Bias in Prediction (MITE)	28
3.7.2	Experiment II: Modality Bias in Attention: MIDAS	29
3.7.3	Experiment III: Calibration with Attention Attribution	32
3.7.4	Discussion	33
3.7.5	Limitations	34
3.8	Conclusion	35
Chapter 4	Formatting Bias within a VLM for VQA	37
4.1	Terminology	37
4.2	Chapter Summary	38
4.3	Introduction	39
4.4	Related Work	40
4.4.1	Towards Formatting Bias Mitigation	40
4.4.2	In-Context Learning (ICL)	41
4.4.3	Out-of-Distribution (OoD) Generalization	41
4.5	Methodology	42
4.5.1	Proposed Methodology Overview	42
4.5.2	Preliminaries	42
4.5.3	Representational Shift Theory (RST)	45

4.5.4	Hypotheses on OoD Generalization	46
4.6	Experimental Settings	47
4.6.1	Models	47
4.6.2	Datasets	47
4.6.3	One-Shot Example Selection	48
4.6.4	Prompt Templates	48
4.6.5	Metrics and Representations	49
4.6.6	On-the-Fly OoD ICL Evaluation	49
4.7	Results & Discussion	50
4.7.1	Experiment I: Multi-image Multi-turn OoD	50
4.7.2	Experiment II: Single-image Multi-turn OoD	53
4.7.3	Discussion	54
4.7.4	Limitations	56
4.8	Conclusion	57
Chapter 5 Conclusion		63
5.1	Are the two studies disjoint?	63
5.2	Contribution Summary	64
5.3	Limitations	64
5.4	Future Work	65
5.5	Concluding Remarks	66
Appendices		67
Appendix A Appendix for Hateful Meme Detection		67
A.1	Formal Relation between miATE and MIDAS	67
A.2	Further Exploration for Local Explainability	68
A.3	Breakdown of Attention Attribution Score	68
Appendix B Appendix for Out-of-Distribution Generalization		75
B.1	Formalization of Representational Shift Theory	75
B.1.1	Representational Shift	75
B.1.2	Formatting Bias Mitigation	75
B.1.3	OoD Generalization as Formatting Bias Mitigation	76
B.2	Other Formalization	77
B.2.1	Mixed Effect Model	77
B.2.2	Representational Analysis	77
B.3	Implementation Details	78
B.4	Additional Results	78
B.4.1	Performance Summary of LLaVA-1.5	78
B.4.2	Additional Analysis on Qualitative Results	78

B.4.3	Mixed Effect of Semantics and Formatting Bias: An Arbitrary Argument?	79
B.4.4	High-Level Analysis on Mixed Effect	80
B.4.5	Preliminary ID Analysis: InternVL	80
B.5	Other Considerations	82
B.5.1	Potential Risks	82
B.5.2	Ethical Considerations	82
B.5.3	AI Assistant Usage	82
B.5.4	License and Usage of Scientific Artifacts	82
B.5.5	Documentation Of Artifacts	82
References		85
Publications		103

Chapter 1

Introduction

1.1 Terminology

In Table 1.1, I list the definitions of the ambiguous terms used in this dissertation.

Table 1.1: Definitions of key terms.

Term	Definition
Bias	Systematic deviation in model predictions from the true underlying distribution.
Presentation Bias	Bias arising from the way information is presented, including modality and formatting influences. Note that the previous machine learning literature focused on the presentation bias in human-machine interaction [5], such as the bias of popular items within an e-commerce website on the recommendation system running on that website [6]. Although the topic of interest may look diverse, I argue that the bias studied in this thesis is an instantiation of the identical concept; I study the bias of presentation on <i>machine</i> , while the other studies emphasized the bias on <i>human</i> .
Modality Bias	Bias due to the choice of input modality (e.g. image vs. text) affecting model behavior.
Formatting Bias	Bias introduced by specific formatting choices (e.g. punctuations and special characters) that alter semantic interpretation.
Alignment	Degree to which a model's behavior faithfully matches intended human objectives and constraints.

continued on next page

Table 1.1 – *continued from previous page*

Term	Definition
Causal Inference	Methodological framework (e.g. structural causal models, do-calculus) for identifying causal relationships via interventions.
Interpretability	Set of techniques for explaining a model’s internal processes and decision rationale.
Attention Attribution	Technique assigning importance scores to input features based on a model’s attention weights [3].
XAI (Explainable AI)	Broad class of methods designed to provide transparent, human-understandable explanations of model decisions.
Semantic Information (or Semantics)	Meaning-related content and relationships in data, distinct from surface-level statistical patterns.

1.2 Background and Motivation

Parallel to the mass production of the graphical processing units, deep learning utilizes its parallel computation capability to show up as a social phenomenon (e.g. Nobel Prize of Chemistry 2024 [7]¹). Originating in an outstanding capability in computer vision [8], its applications are widespread across various scientific and social domains. In Natural Language Processing (NLP), following the success of Bidirectional Encoder Representations from Transformers (BERT) [9] and Generative Pre-trained Transformer (GPT) [10–12], Transformer [13] becomes the de facto standard architecture of Language Model (LM)s [14]. Recently, Transformer’s application is extended to Vision-and-Language Model (VLM)s designed to solve Vision-and-Language (VL) tasks, consisting of a combination of relevant images and texts [15, 16]. In order to prevent the potential risks caused by VLMs, the alignment between a human objective and the VLMs is crucial. To achieve this human-VLM alignment, a key challenge is to detect and mitigate the various *biases*, or the factors irrelevant to the task that affect the outcome, in a VLM’s decision making.

Previous studies tackled two types of biases. The first type of bias is *social bias*: the bias towards a specific protected group of people in socially impactful decisions, for which the famous example is the overestimated recidivism risk for black people [17]. Although the social bias is critical

¹<https://www.nobelprize.org/prizes/chemistry/2024>

for achieving the social good, I focus on the second type of bias that could affect broader decisions—what I term here as *presentation bias* (see Figure 1.1 for an overview). In contrast to the social bias, or the effect of *what* is presented in the text and/or image, the presentation bias can be defined as the bias caused by *how* the content is presented (whether it is in the image or in the text, how the input is formatted, etc.; also refer to Table 1.1). By definition, this could affect *all* the decisions, such as answering a general question (e.g. "Is there a car in the image") about the presented image (e.g. an image about a car, or something else). In this dissertation, I discuss two main subtypes of presentation biases: the *modality bias* and the *formatting bias*.

The first type of presentation biases, *modality bias*, is defined as the over-reliance to a modality, ignoring others. In VL domain, both the bias towards the image and towards the text are reported. An interesting and socially important problem for studying modality bias is *Hateful Meme Detection (HMD)*; Hateful Meme (HM) is the hateful message expressed in a combination of text and image, and HMD is the task of detecting this hateful message. A notable benchmark in this topic is the Hateful Memes Challenge (HMC) dataset [18]. First, this dataset is tailored so that the specific *combination* of an image and a text, not a single modality, expresses the hateful message. Second, to challenge the reasoning behind the detection, its counterpart *confounders* in which either one of the two modalities (text *or* image) is replaced so that the replacement flips the label to benign (i.e., not hateful). Together, the HMC dataset is designed to challenge the VLMs' reasoning ability in the presence of the modality bias.

The second type of presentation biases is *formatting bias*. This bias is defined as the bias of the input format affecting the performance regardless of the semantic information conveyed. A well-known example is prompt sensitivity [19, 20], where the Large Language Model (LLM)s and VLMs are affected by the details of the instruction prompt (e.g. paraphrasing of a word and absence/presence of a special character). Detecting and mitigating this bias is particularly important for conversational agents, since the choice of formatting (e.g. of words and characters² are both appropriate for the intention.) is diverse. They solve most problems as Visual Question Answering (VQA) tasks: answering questions based on VL inputs. For example, a web service provider may ask a VLM for a question "*Is this meme hateful or benign?*" to solve an HMD task. Therefore, addressing formatting bias in VLMs for VQA tasks is critical. In particular, one of

²For example, to ask an agent to answer in one word, both "*Answer in a single word.*" and "*You should reply in a single-word answer like 'Yes'.*"

the most challenging formatting biases in VQA is *the number of turns and images*. In particular, if a turn of the conversation results in insufficient information for answering the question, the VLM may need to refer to the previous turns. However, most of the open-sourced VL datasets consist of *single-image single-turn resources*, which makes it difficult for the VLMs to fully leverage their ICL capacity.

To detect and mitigate the presentation biases, the methodologies of the previous studies can fall into *debiased training*³ and *causal inference*. The first approach, debiased training, focuses on the bias in the training data and use the unbiased subset for training (e.g. [21]) or design an optimizer weighted over the subsets (e.g. [22]). While this approach results in an inherently debiased VLMs, applying these approaches to an *existing* pretrained model is challenging. The latter approach, causal inference, addresses this problem. For example, [23] proposed a VQA benchmark and a dynamic pipeline for enhancing a VLM’s utility of both image and textual cues.

In summary, one of the problems critical for the VLM alignment is *how we can detect and mitigate the presentation biases of the trained VLMs in important problems in society and/or VL domain, such as the modality bias in HMD and the formatting bias in VQA*. My motivation in this dissertation is to contribute to this area by *science and engineering combined*: how VLMs work when solving these problems, and how the understanding of those inner workings can benefit the original tasks.

1.3 Problem Statement

To assess interaction of the presentation biases and the inner workings of the trained VLMs, an intuitive approach is to utilize the methodologies of eXplainable Artificial Intelligence (XAI) and interpretability. To this end, there exist a few research gaps.

1. *Detecting the presentation biases in a trained VLM’s inner workings.* Previous efforts on presentation bias detection have largely focused on the outcome (e.g. VLM’s predictions or responses). For example, [24] introduced a dual-model framework, a model trained solely with the text and the other solely with the image, for curating counterfactual labels for calibrating the biased VQA models. However, it may be *right for the wrong reason*; Debiasing a model solely based on its outcome

³In this research, the term *training* includes the processes which require the large-scale gradient computation of a VLM, such as pertaining, supervised fine-tuning, and reinforcement learning from human feedback.

poses a potential risk of misinterpretation and less generalization.

2. *Debiasing a VLM’s decisions upon insights about its inner workings.*

To alleviate the presentation biases in LLMs and VLMs, previous works proposed various outcome-based methods, such as a bias-aware correction of the prediction (e.g. [25, 26]) and front-door adjustments in the model’s reasoning process [27]. Those outcomes, however, may not fully reflect the model’s capacity; For example, [28] reported that the LLMs’ answers on factual knowledge questions may not be able to answer a question for which the required knowledge is fully encoded in latent space.

1.4 Research Objective

The goal of this research is to build a framework for interpreting how the presentation biases affect the VLMs’ inner workings, and assess how this interpretation could contribute for building a better problem-solving pipeline for HMD and VQA.

1.5 Dissertation Outline

Toward the research objective stated in §1.4, the rest of this dissertation is four-fold. First, I showcase previous works related to this dissertation (Chapter 2). Next, I show the details of how the interpretability methods could be used for detecting and alleviating presentation biases; the modality bias in BERT-based VLMs for HMD (Chapter 3) and the formatting bias in generative VLMs for HMD and VQA (Chapter 4). Finally, I finish this dissertation with concluding remarks (Chapter 5).

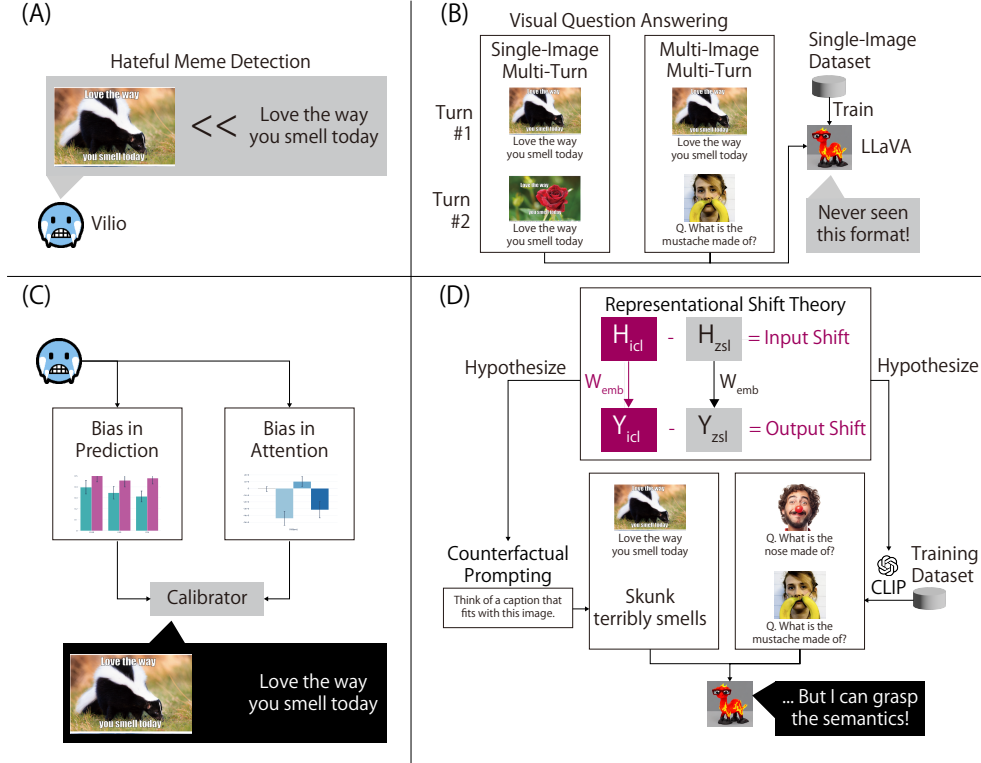


Figure 1.1: Schematic overview of the entire dissertation. (A,B) Presentation biases studied in this dissertation. (A) Modality bias in Hateful Meme Detection (HMD; Chapter 3). A State-of-the-Art (SotA) HMD model Vilio [1] suffers from the bias towards text modality, potentially due to the shortage of multimodal training. (B) Formatting bias in Visual Question Answering (VQA) (Chapter 4). Two variants of LLaVA [2], which exhibit SotA performance in VQA and are trained solely with single-image and single-turn vision-and-language resources, struggle with multi-image and/or multi-turn VQA queries. (C,D) High-level overview of the proposed methodologies. (C) Approach for modality bias in HMD. We propose two metrics for quantifying the modality biases in a model’s predictions and in attention attribution [3]. Given the insights about the biases within the model, we calibrate the prediction with attention attribution for debiasing the base predictor. (D) Approach for formatting bias in VQA. First, we propose Representation Shift Theory (RST), a novel theory about ICL for formatting bias mitigation. Next, based on RST, we empirically validate two hypotheses for the bias mitigation: 1) prompt engineering can contribute to the performance gain in a text-predominant task (validated by counterfactual prompting for HMD), and 2) careful VQA example selection can lead to better performance when the model struggles with the task in the zero-shot setting (validated by CLIP-based example selection for VQA).

Chapter 2

Related Work

2.1 Outline

In this chapter, I provide an overview of the works related to this dissertation. First, I present a summary of recent advancements of LMs, LLMs, and VLMs (§2.2). Next, I provide more details on the VL applications, specifically VQA (§2.3) and HMD (§2.4). Then, I discuss the biases in VLMs, focusing on modality bias (§2.5.1) and formatting bias (§2.5.2), as well as their detection and mitigation techniques. Finally, I summarize the relevant interpretability methodologies, including causality-oriented interpretability (§2.6.1), interpretability for bias mitigation (§2.6.2), and applications of interpretability in VLMs (§2.6.3).

2.2 LMs and LLMs

The introduction of the Transformer architecture [13] marked a paradigm shift in sequence modeling by replacing recurrence with self-attention, enabling highly parallelizable training and forming the backbone for LMs; BERT [9] then showed that a bidirectional Transformer encoder trained via masked language modeling could be fine-tuned to achieve State-of-the-Art (SotA) results across diverse NLP benchmarks, while the GPT series [10–12] demonstrated the fluency and versatility of autoregressive Transformer decoders. Building on these foundations, we saw unprecedented scaling and architectural innovation: GPT-3’s 175 billion parameters introduced ICL, a gradient-free way of learning solely based on the input prompt.

Given the success of the unimodal applications in NLP, the research attention has been shifted to their multimodal capabilities [29], centered on VL as a step toward general-purpose agents. Despite these advances, the mechanisms by which such large models integrate and prioritize multimodal signals remain opaque.

2.3 VQA

VQA is the task of automatically answering natural language questions about the contents of an image, requiring models to interpret visual data and understand linguistic queries in tandem. VQA has matured from early Convolutional Neural Network (CNN)–Recurrent Neural Network (RNN) models with bottom-up attention over object proposals [30] to fully Transformer-based, multimodal architectures that learn joint image–text representations (e.g. [31]). The original VQA benchmark [32] spurred methods like Bottom-Up/Top-Down Attention [30] to focus on salient image regions; soon after, bi-modal Transformers such as ViLBERT [31] and LXMERT [33] introduced cross-attention layers between region features and question tokens, yielding large gains in context-sensitive reasoning. The paradigm shifted further when contrastive pre-training at internet scale—exemplified by CLIP [11]—demonstrated that simple image–text matching produces highly transferable embeddings capable of zero-shot VQA. Generative unified models like Simple Visual Language Model (SimVLM) [34] simplified pre-training by treating image patches as prefix tokens in a language-modeling objective, achieving SotA supervised VQA accuracy and strong zero-shot generalization. More recently, few-shot multimodal learners such as Flamingo [35] have extended ICL to interleaved sequences of images and text, enabling open-ended visual reasoning without task-specific fine-tuning. [36] bridges pre-trained vision encoders and frozen language models via a lightweight query transformer. By instruction-tuning such multimodal models on diverse tasks, [37] attained strong zero-shot and fine-tuned performance across many VQA benchmarks. 17-billion-parameter PaLI model [38] reached 84.3% accuracy on the open-domain VQAv2 [39] benchmark. Recent VLMs have also demonstrated unprecedented flexibility in visual reasoning – for example, [40] showed strong few-shot results on tasks ranging from captioning to open-ended VQA without any task-specific fine-tuning. These developments underscore a shift toward general-purpose VQA systems that leverage massive data and parametric knowledge, yielding both high accuracy and the ability to generalize zero-shot to novel domains.

2.4 HMD

HMD is the task of identifying memes that convey hateful content by jointly analyzing the combination of imagery and overlaid text, particularly when neither modality alone suffices to detect the hateful intent. Sparked by the Hateful Memes Challenge [18], this problem requires models to discern

subtle meaning emergent from the image–text pairing. Note that *benign confounders* which have identical images (or captions) but their paired captions (or images), ensure that either one of the two modalities are insufficient for the hateful message. Early Transformer-based fusion baselines performed well on unimodal cues but struggled to distinguish subtly altered benign variants [41]. To address this, various approaches have been proposed, such as model ensemble [1], multi-task learning [42], OCR integration [43], and ICL [44]. More recently, one line of research has enriched the input and reasoning process: for instance, [45] combine OCR text extraction, neutral image captioning, retrieval-augmented inference, and a VQA module into an iterative framework that uncovers subtle hate signals (e.g. sarcasm or coded imagery), outperforming conventional fusion models on the HMC benchmark. Other approaches leverage the latest pre-trained vision–language representations, such as an ensemble of Contrastive Language-Image Pre-Training (CLIP)-based visual features with a BERT text classifier [46]. Despite these advances, generalization beyond the training domain remains problematic; A cross-domain study found that current multimodal classifiers often rely overwhelmingly on the text cue, struggling when meme images differ from training data [47].

2.5 Bias Detection and Mitigation

Historically, the term *bias* referred to the *social* bias: an unfair treatment of specific personal demographics, such as race or gender, or its combination [48]. Therefore, research focus has been on how to detect or mitigate such biases inherent in machine learning models [49, 50].

In parallel to the explosive growth of LMs, LLMs and VLMs, another source of biases has garnered attention: data structure. Here I focus on two major such biases, namely the modality bias (§2.5.1) and the formatting bias (§2.5.2).

2.5.1 Modality Bias

Modality bias is defined as an overreliance on a single modality, either vision or language in VL field, at the expense of integrating complementary information from the other channel, leading to brittle or misleading predictions. In the context of HMD, this manifests when models focus almost exclusively on the overlaid text or the image background, failing to capture the hateful intent that arises only from their combination. Early diagnostic works such as Counterfactual VQA probes language bias in VQA by swapping question

variants and measuring performance shifts [51]. Similarly in HMD, ensemble and multi-task architectures initially demonstrated that unimodal baselines often rival full fusion models, underscoring the need for bias detection.

To mitigate modality bias, researchers have explored several strategies. [39] elevated the role of image understanding in VQA by enforcing object-level attention, which inspired analogous approaches in HMD to strengthen visual feature extraction [1, 18, 42]. Additionally, [52] proposed a collaborative modality fusion mechanism that dynamically balances visual and textual embeddings based on mutual agreement, while [23] formalized unimodal bias quantification through a regularization term that penalizes overreliance on either modality. More recently, [53] leveraged noise-aware prefix tuning to recalibrate modality contributions at inference, achieving robust performance under textual or visual confounders. Complementary data-centric methods, such as contrastive [54] and counterfactual augmentations [51], further force models to attend to semantically critical differences, reducing shortcut learning.

However, these methods typically operate at the model output or data level, without revealing internal causal pathways. We bridge this gap by proposing two metrics, one for a causal effect measure on model outputs and the other on an attention-based attribution score, to pinpoint where shortcuts occur and to enable targeted debiasing via a downstream re-ranker.

2.5.2 Formatting Bias

Formatting bias refers to the influence of input presentation on model outputs independent of the underlying semantic content. Here input presentation includes, but not limited to, prompt phrasing, token ordering, or punctuation. For example, [19] shows that small paraphrases or the presence/absence of special characters can drastically alter factual knowledge extraction in LLMs. This instability undermines reliability in both conversational agents and VL tasks, where inputs are often user-generated.

Mitigation strategies for formatting bias span prompt engineering, training-time robustness, and architectural adjustments. Prompt diversification—which generates multiple semantically equivalent variants—and self-ensemble inference have been shown to stabilize outputs. For example, [55] identified position bias in VLMs by systematically permuting token orders and introduced positional decorrelation layers to neutralize formatting artifacts. Additionally, adversarial prompt training—where models are exposed to worst-case paraphrases at training time—has proven effective in reducing output variance across formatting changes for LLMs [19]. Together, these methods move toward LLMs and VLMs that are robust not only to what is

said, but also to how it is phrased.

Building on these works, we propose a theory-driven framework for diagnosing and correcting non-semantic input artifacts in VLMs for VQA. We quantify the changes in latent space between zero-shot and in-context embeddings, introduce two ICL strategies that exploit this change, and demonstrate their effectiveness on six standard VQA datasets and the HMC benchmark.

2.6 Interpretability

Interpretability methods provide crucial insights into the inner workings of LLMs and VLMs, revealing how presentation biases arise and guiding effective debiasing strategies. Here I focus on three research directions. First, to understand *why* the bias is present in specific input, or what *caused* the bias, a line of work introduced the notion of causality in interpretability (§2.6.1). Second, towards the real-world unbiased language model applications, the researchers also proposed interpretable debiasing methodologies (§2.6.2). Finally, I summarize the existing applications of interpretability to VLMs (§2.6.3).

2.6.1 Causally Interpreting LLMs

Recent advances in causally interpreting Transformer internals is centered on Mechanistic Interpretability (MI). In this field, the researchers dissect LMs using tools from mechanistic analysis and causal inference. They decompose *circuits* in LMs by instrumenting individual attention heads and feed-forward network blocks [56,57]. For example, the logit lens [58] projects intermediate hidden states into the vocabulary space, quantifying each component’s contribution to final logits, which exposes shortcut pathways and latent bias vectors.

These methods enable targeted interventions within network components to trace the causal chains influencing model outputs. However, they often stop short of connecting these causal insights back to the models’ optimization trajectories; in this dissertation, we address that by introducing causality-oriented metrics and hypothesis-driven ICL strategies.

2.6.2 Interpretability for Debiasing LLMs

Building on these causal insights, [59] leverages MI to identify the Transformer components that systematically skew predictions toward spurious

labels; masking these components at inference yields substantial gains in both accuracy and robustness under prompt perturbations. Complementary calibration techniques, such as contextual calibration [60] and domain-context/prototypical calibration [61, 62], adjust model output probabilities based on estimated bias magnitudes, aligning predictions with desired distributions without modifying model weights.

In another line of work, Representation Engineering (RepE) [63] has recently emerged as a top-down paradigm for interpretability and bias control in LLMs, focusing on identifying linear concept vectors in latent space by contrasting model activations instead of dissecting individual neurons. This approach enables researchers to read and manipulate high-level internal representations of attributes like honesty, toxicity, or demographic features. For example, probing the activations of a chat model can reveal persistent social biases (e.g. gender bias remains detectable even after training [64]), indicating that such biases are encoded along specific directions in the representation space. To mitigate unwanted biases, recent work leverages linear interventions on these representations. Building on these ideas, Minimally Modified Counterfactuals (MiMiC) applies an affine transformation that matches the mean and covariance of hidden representations for different demographic inputs, so that protected attributes no longer correspond to separable clusters [65]. Such representation-space counterfactuals effectively remove the bias information and have been shown to reduce toxic or biased outputs without any model retraining. This line of research complements MI by treating an LLM’s hidden states as a linear algebraic space of concepts that can be read and edited, offering a promising strategy to uncover latent biases and directly manipulate the model’s representations to achieve debiasing and safer performance.

While these works pinpoint where models focus, they rarely translate those findings into systematic mitigation strategies. Chapter 3 remedies this by integrating attention attributions with a bias-aware re-ranking stage, yielding an end-to-end pipeline that both diagnoses and corrects modality shortcuts in the HM case study. Moreover, Chapter 4 breaks new ground with Representational Shift Theory (RST), leveraging in-context demonstrations to quantify representational shifts and inform prompt engineering tactics that counteract formatting artifacts in VQA models.

2.6.3 VLM Applications

In VLMs, interpretability not only explains decisions but also drives bias mitigation. Attention heatmaps and gradient-based saliency maps (e.g. [66]) visualize which image regions and text tokens the model relies on,

uncovering modality shortcuts. Prototype-driven modules—such as Dynamic Clue Bottlenecks [67]—interpose an information bottleneck that filters out presentation artifacts (e.g. fixed layouts or overlaid text styles), enabling targeted correction of modality and formatting biases in tasks like VQA and HMD.

Although these methods offer descriptive insights, they lack a causal or theoretical underpinnings to guide robust bias mitigation. Chapter 3 addresses this by embedding our causal metrics into the HM detection pipeline. In turn, Chapter 4 proposes an analytical framework across standard VQA benchmarks and the HMC benchmark: by measuring how in-context examples reshape latent representations, we derive prompt-level interventions that bolster model performance under the strong formatting biases. Together, these contributions demonstrate a cohesive, causally informed interpretability toolkit for VLM applications.

Chapter 3

Modality Bias within Hateful Meme Detectors

3.1 Terminology

Table 3.1: Key terms used in Chapter 3.

Term	Definition / Reference
Hateful Memes Challenge (HMC)	Benchmark of paired hateful vs. benign confounder memes [18].
Text-to-Text (T2T) / Image-to-Image (I2I) tasks	Evaluation regimes where one modality is fixed and the other is swapped to measure formatting bias; see §3.5.3.
Causal Intersectionality	Extension of TE to interacting binary factors (e.g. text-hateful vs. image-hateful); see §3.5.1.2.

3.2 Chapter Summary

In this chapter, we examine modality shortcuts in SotA hateful meme detectors through a unified framework driven by causality and interpretability. We begin by highlighting how models can overly rely on text or image alone, leading to spurious *hate* predictions when the other modality carries the true signal (§3.3). After surveying existing multimodal bias metrics and attribution methods (§3.4), we introduce two novel tools (§3.5): *Multimodal Intersectional Treatment Effect (MITE)*, a causal metric that measures how interventions on text or image affect model outputs, and *Modality Interaction Disentangled Attribution Score (MIDAS)*, an attention-based score that disentangles within-text, within-image, and cross-modal interactions in the final attention layer. We then detail our experimental setup on the HMC

dataset (§3.6): dataset overview, controlled intervention protocols, and feature extraction from BERT-based models that we analyze (Oscar, UNITER, and VisualBERT). In Experiment I (§3.7.1), MITE reveals pronounced text biases, especially in VisualBERT, by quantifying shifts in *hate* probability under modality swaps. In Experiment II (§3.7.2), MIDAS uncovers internal attention imbalances at both global and local levels, showing how models distribute attribution across modalities. Finally, in Experiment III (§3.7.3), we demonstrate that injecting MIDAS features into a LightGBM [68] re-ranker consistently improved macro-AUC across multiple sub-models. We conclude by summarizing our contributions—causal intervention metrics, attention-based interpretability, comprehensive bias analysis, and bias-aware calibration—and by outlining future directions for extending our framework to other multimodal tasks (§3.8).

3.3 Introduction

Hateful memes present a unique challenge in multimodal understanding: neither the image nor the text alone fully conveys hateful intent, yet current detectors often rely disproportionately on one modality, leading to spurious predictions. This chapter asks: *How can we systematically detect when a model shortcuts on text- or image-only cues, and how can we mitigate these modality biases to build more robust hateful meme detectors?*

Prior efforts in HMD have predominantly focused on predictive performance, with recent studies exploring causal and attribution-based analyses to uncover biases. Causal intersectionality frameworks measure synergistic demographic effects in multimodal settings, while explainability techniques split attention into text-only and image-only categories. Adversarial robustness work examines detector failures under perturbations and proposes fine-tuning defenses. However, these approaches either offer only coarse, two-way attributions, neglect attention’s cross-modal interactions, or treat bias purely as an adversarial phenomenon without integrating diagnostic indicators into the detection pipeline. As a result, no unified methodology exists to both pinpoint and correct modality shortcuts in SotA VLMs.

We introduce a two-fold framework combining causal intervention and fine-grained interpretability. First, for three VLMs (Oscar, UNITER, and VisualBERT), **MITE** quantifies output-level modality bias by measuring the change in hate-probability when text and image jointly convey hate versus when each is isolated. Second, **MIDAS** partitions last-layer attention attributions into within-text, within-image, and cross-modal interactions to reveal where, inside the network, shortcuts occur. Finally, we inject these

bias indicators, MITE and the three MIDAS features, into a lightweight LightGBM re-ranker, effectively down-weighting unimodal shortcuts and improving macro-AUC for the models with stronger shortcuts.

Our contribution is summarized as:

- **Causal Metric for Modality Bias (MITE).** A formal intersectional effect measure that captures reliance on joint text–image signals versus single-modality confounders.
- **Attention-Based Attribution for Modality Bias (MIDAS).** An attention attribution metric disentangling within-text, within-image, and cross-modal attention patterns to diagnose internal shortcuts.
- **Comprehensive Bias Analysis.** Empirical evaluation of MITE and MIDAS across three SotA VLMs—Oscar, UNITER, and VisualBERT—revealing pronounced textual shortcuts, especially in VisualBERT.
- **Bias-Aware Calibration.** A LightGBM re-ranking strategy that leverages MITE and MIDAS features to mitigate modality biases and boost classification performance.
- **Interpretability Meets Bias Mitigation.** A unified pipeline that not only diagnoses but also corrects modality shortcuts, advocating for causal intersectionality analyses in multimodal model evaluation.

3.4 Related Work

3.4.1 Intersectionality and Causal Analyses

Although much hateful-meme research focuses on purely predictive performance, recent work has begun to examine how multiple, overlapping demographic attributes (*intersectionality*) affect a model’s internal behavior. [48] formalize a causal framework for measuring how intersecting factors (e.g. gender \times race) jointly influence outcomes in multimodal settings. They show that simple additive biases can obscure deeper, synergistic intersectional effects. Our chapter draws on these ideas by defining MITE to quantify how much an HM detector relies on joint (text and image) cues instead of one modality alone—an intersectional effect in the representation space.

3.4.2 Explainability for Hateful Meme Models

[4] examines *what* VLMs actually learn when fine-tuned on the HMC dataset [18]. They introduce attention attribution score (attr)-based analysis, an attribution technique that splits attention into two categories (text-only and image-only), and then measures how these changes occur between a hateful

meme and its benign *confounder* variant. Their error analysis shows that VisualBERT and VilBERT [31] often ground spurious keyword–image pairs (e.g. *dishwasher* to image of a woman) or maintain over-sensitive slur-to-image alignments (e.g. *goat* to bounding boxes of Muslim men), leading to false positives. We build upon these findings by (1) introducing a third category (cross-modal) for a better interpretation, and (2) using these bias indicators to *mitigate* shortcuts via LightGBM calibration.

3.4.3 Adversarial Robustness and Countermeasures

Concurrently with this work, [47] thoroughly studied how current VLM-based detectors break under *human-induced*, *partial-knowledge* adversarial attacks, such as adding noisy text tokens (“*LOVE*”) or injecting salt-and-pepper noise into the image region. They demonstrated that even small perturbations cause a sharp drop (up to 10%) in macro-F1, and then propose two complementary defenses: (i) a contrastive-learning augmentation at embedding level, and (ii) an adversarial fine-tuning. Their ensemble of these two methods restores most of the lost performance on high-resolution subsets like HMC, though low-resolution datasets such as MAMI [69] remain challenging. Using MITE and MIDAS to diagnose internal biases, our work diverges by focusing not just on *whether* the model is robust, but also *why* it shortcuts to modalities in the first place, rather than only testing external attacks.

3.4.4 Datasets and Multimodal Architectures

Several datasets have been constructed to capture different flavors of hate and offensiveness in images-plus-text:

- **HMC** [18] provides 10 K crowdsourced image–text pairs specifically labeled for hate, with carefully balanced *benign confounder* variants (e.g. same image with benign text, or vice versa) to force cross-modal reasoning.
- **HarMeme** [70] collects COVID-19-influenced hateful memes from social media; text is automatically extracted via OCR.
- **MAMI** [69] focuses on misogynistic memes; they re-label misogynistic vs. non-misogynistic as hateful vs. non-hateful for consistency.

Most top-performing detection methods rely on large pretrained VLMs (e.g. VisualBERT, VilBERT, Oscar, and UNITER), which fuse region-level image features (e.g. by Faster-RCNN [71]) with token embeddings from BERT [9] and then perform late-fusion classification. We evaluate three canonical

architectures (Oscar, UNITER, and VisualBERT), and show that two of them (VisualBERT and Oscar) exhibit modality shortcuts that can be both detected via MITE and MIDAS, and partially mitigated.

3.4.5 Broader Bias and Debiasing Approaches

Beyond hateful memes, recent studies extend modality-bias research beyond the HMC dataset and the VLM fusion architectures evaluated in this chapter. For example, [53] focused on debiasing large VLMs via preference optimization; [72] introduced a unified, inference-time feature-pruning strategy that applies to any frozen VLM component; and [73] demonstrate how auxiliary bias estimation in VQA can guide adaptive loss reweighting. All of these methods share a common goal: to reduce the model’s over-reliance on a single modality, whether in classification, generation, or question answering, without significantly degrading overall performance. Our work on MITE and MIDAS complements these advances by providing diagnostic tools to quantify and interpret modality shortcuts in HMD models and by proposing a bias-aware calibration that can be viewed as a lightweight analogue to the more heavy-weight methods above (e.g. [72]’s feature imputation or [53]’s dataset and optimization changes), tailored specifically to the HMD task.

3.5 Methodology

In this section, we present our pipeline for (i) detecting modality bias in HM detectors and (ii) mitigating its harmful effects. We begin by introducing the causal-inference preliminaries (Treatment Effect (TE) and intersectionality) and the attr mechanism (§3.5.1). We then give an overview of the proposed methodology, illustrated in Figure 3.1. Finally, we formalize the intersectional effect of a meme over the VLM’s prediction (MITE; §3.5.3) and over the attention (MIDAS; §3.5.4).

3.5.1 Preliminaries

Before describing our proposed metrics, we review three key building blocks: (a) the standard TE in causal inference, (b) its extension to causal intersectionality, and (c) attention-based attributions for Transformer models.

3.5.1.1 Treatment Effect (TE)

In causal inference, the TE measures the expected difference in an outcome when a binary *treatment* variable flips from 0 to 1. Formally, for a binary treatment $T \in \{0, 1\}$ and a probabilistic outcome $\theta_{T=i}$ given the treatment T set to value i , the TE is

$$\text{TE} = \theta_{T=1} - \theta_{T=0} \quad (3.1)$$

This quantity captures how much the presence of *treatment* changes the outcome. In our context, we will treat each modality (text or image) as a *treatment* whose presence or absence can jointly affect the model’s predicted probability of a meme being hateful.

3.5.1.2 Causal Intersectionality

Standard TE aggregates only over a single binary factor. However, when two or more binary factors (e.g. “*text is hateful*” and “*image is hateful*”) interact non-additively, we require a notion of *intersectional effect*. Following prior work [48], we define the causal intersectionality between two binary variables $D_1, D_2 \in \{0, 1\}$ (e.g. text-hateful indicator and image-hateful indicator) on outcome $\theta_{D_1=i_1, D_2=i_2}$ by

$$\theta_{D_1=1, D_2=1} \neq \theta_{D_1=1, D_2=0} + \theta_{D_1=0, D_2=1} \quad (3.2)$$

In other words, when the joint effect of $(D_1 = 1, D_2 = 1)$ differs from the sum of each single-factor effect, there is a nontrivial *intersectional* influence. We will reframe this concept so that D_1 corresponds to *text indicates hate* (binary) and D_2 corresponds to *image indicates hate* (binary).

3.5.1.3 Attention Attribution Score

To diagnose *where* a Transformer model is attending on its cross-modal reasoning (text-only shortcuts, image-only shortcuts, or true text–image synergy), we use an attr metric. Let

$$A \in \mathbb{R}^{L \times L} \quad (3.3)$$

denote an attention-weight tensor with sequence length L (for simplicity, flattening all heads and positions into one two-dimensional matrix). Given a scalar model output $\theta(A)$ for which we regard the predicted hatefulness θ as a function of the attention matrix A , the attribution of A can be written as

$$\text{attr}(A) = A \odot \int_{\alpha=0}^1 \frac{\partial \theta(\alpha A)}{\partial A} d\alpha \quad (3.4)$$

where \odot is element-wise multiplication and α is a hyperparameter indicating the number of steps in integrated gradient (the integral term, a step-wise gradient between a model’s component of interest and the prediction calculated as an attribution of that component [74]) computation. Concretely, $\text{attr}(A)$ has the same shape as A , and each entry $\text{attr}(A)_{i,j}$ measures how much that particular attention weight contributes to the final output θ . We will compute this at the last attention layer of our VLMs and then aggregate by modality-interaction type (within-text, within-image, or cross-modal).

3.5.2 Proposed Methodology Overview

Figure 3.1 visualizes our overall pipeline. After we extract paired samples (T_1, I_1) (a truly hateful meme) and its *benign confounders* (T_1, I_0) (image replaced) and (T_0, I_1) (text replaced) from the original datasets, our pipeline comprises three major steps:

1. **Causal Effect Estimation via MITE.** We feed each hateful-benign pair into the frozen VLM to obtain probabilities θ . MITE is then computed as the prediction difference between a hateful sample and its counterpart confounders (Eq. 3.6). A large positive MITE indicates that the model pays attention to *joint* cross-modal cues, whereas a near-zero or negative MITE implies that the model is shortcutting on a single modality (text-only or image-only).
2. **Modality Attribution Analysis via MIDAS.** For each hateful-benign triplet (T_1, I_1) , (T_1, I_0) , (T_0, I_1) , we extract the last-layer attention tensor A and compute $\text{attr}(A)$. We then split attr into three interaction categories *it*:
 - **Within-Text:** attention between two text-token positions,
 - **Within-Image:** attention between two image-region positions,
 - **Cross-Modal:** attention between a text token and an image region (in either direction).

Let MIDAS_{it} be the averaged difference of attention attribution over the interaction type *it*. Intuitively, $\text{MIDAS}_{\text{text}}$ measures how much *Within-Text attention* increases when *both* modalities jointly signal hate, relative to either modality alone; likewise for $\text{MIDAS}_{\text{image}}$ and $\text{MIDAS}_{\text{cross}}$. If $\text{MIDAS}_{\text{text}}$ is large and the others are small, the model is internally focusing on textual shortcuts regardless of which modality truly determines hate.

3. **Bias-Aware Calibration.** Having computed the predicted hatefulness θ , attention attribution for each interaction type ($\text{attr}_{\text{text}}$, $\text{attr}_{\text{image}}$, and $\text{attr}_{\text{cross}}$) for each input (T, I) , we train a simple

LightGBM classifier that takes the four-dimensional feature vector $[\theta, \text{attr}_{\text{text}}, \text{attr}_{\text{image}}, \text{attr}_{\text{cross}}]$ and outputs a debiased prediction of the hate label. Empirically, this *bias-aware re-ranking* corrects modality-shortcut errors (especially in VisualBERT, which tends to over-rely on textual cues), indicated by the AUROC gain over the θ -only baseline.

3.5.3 Multimodal Intersectional Treatment Effect (MITE)

In classical causal inference, the TE addresses a single binary intervention (Eq. 3.1). Here, we treat *text is hateful or benign* (T) and *image is hateful or benign* (I) as two binary factors whose joint effect may be more than the sum of their individual contributions (Eq. 3.2). Concretely, suppose we have:

- (T_1, I_1) : an original hateful meme (both text and image jointly convey hate),
- (T_1, I_0) : a *text-benign* confounder (text alone is hateful; image replaced so that it no longer signals hate),
- (T_0, I_1) : an *image-benign* confounder (image alone is hateful; text replaced).

Let $\theta_{T=T_i, I=I_i}$ be the predicted hatefulness generated by any pretrained VLM (e.g. Oscar, UNITER, or VisualBERT), and write

$$\theta_{T=T_1, I=I_1} \neq \theta_{T=T_1, I=I_0} + \theta_{T=T_0, I=I_1} \quad (3.5)$$

as defined by the intersectional effect condition in Eq. 3.2, the meme’s joint effect exceeds the sum of its individual effects. Hereafter we use $\theta_{T=T_1, I=I_0}$ and θ_{T_1, I_0} interchangeably. Then, following the standard definition of TE (Eq. 3.1), we define the MITE as

$$\text{MITE}(\theta; T_1, I_1) = \theta_{T_1, I_1} - (\theta_{T_1, I_0} + \theta_{T_0, I_1}) \quad (3.6)$$

Intuitively:

- If MITE is large and positive, it means that the probability of hate jumps *only when both* text and image jointly indicate hate. Thus, the model that generates the predicted hatefulness θ is correctly relying on cross-modal synergy.
- If MITE is near zero (or negative), it implies that the model does not *need* both modalities to detect hate (i.e., shortcutting on a single modality).

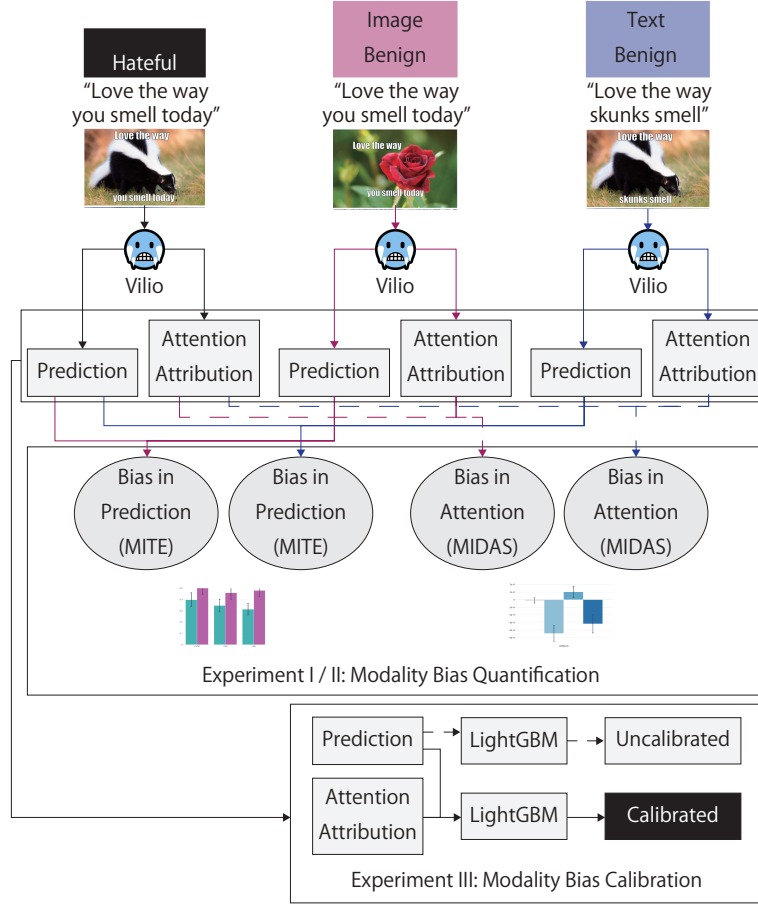


Figure 3.1: Schematic overview of the proposed pipeline. First, a hateful meme and its benign confounder counterpart are given to Vilio for extracting its predicted hatefulness and the attention attribution (top). The confounders are divided into image benign (the image is replaced from the original hateful meme so that it flips the label to benign) and text benign (the text is replaced, flipping the label). In Experiments I and II (middle), we quantify modality bias as the prediction difference between a hateful example and its benign confounder counterpart (MITE; Experiment I) and as the attention attribution difference (MIDAS; Experiment II). To see if attention attribution is effective for calibrating the prediction upon modality bias or not, we then build a LightGBM re-ranker (bottom), solely with the prediction or with the prediction and the attention attribution.

3.5.4 Modality Interaction Disentangled Attribution Scores (MIDAS)

While MITE diagnoses *output-level* shortcuts, it does not reveal *where* inside θ this shortcut occurs. To trace these biases into the attention layers, we

define MIDAS as follows:

1. **Extract the Last-Layer Attention Tensor.** For a given input pair (T, I) , run θ forward and obtain the final attention weight tensor $A \in \mathbb{R}^{L \times L}$ where L is the total number of *tokens* (text tokens or visual regions). We exclude the special tokens [CLS] and [SEP] from the analysis.
2. **Compute attr.** Compute $\text{attr}^{(T,I)}$ (Eq. 3.4). Each entry $\text{attr}_{i,j}^{(T,I)}$ quantifies how much the attention weight from position i to j contributes to the prediction θ .
3. **Partition by Interaction Type.** Index all pairs (i, j) into three disjoint sets:

$$\begin{aligned}\mathcal{P}_{\text{text}} &= \{(i, j) : i, j \text{ are both text tokens}\}, \\ \mathcal{P}_{\text{image}} &= \{(i, j) : i, j \text{ are both image regions}\}, \\ \mathcal{P}_{\text{cross}} &= \{(i, j) : i \text{ text}, j \text{ image}\} \cup \{(i, j) : i \text{ image}, j \text{ text}\}.\end{aligned}$$

4. **Compute MIDAS_{it}.** For each interaction type $it \in \{\text{text}, \text{image}, \text{cross}\}$, define

$$\text{MIDAS}_{it} = \frac{1}{|\mathcal{P}_t|} \sum_{(i,j) \in \mathcal{P}_t} \left[\text{attr}_{i,j}^{(T_1, I_1)} - (\text{attr}_{i,j}^{(T_1, I_0)} + \text{attr}_{i,j}^{(T_0, I_1)}) \right] \quad (3.7)$$

In words, MIDAS_{it} measures how much *attention mass* on interaction type it *increases* when both modalities are truly hateful, compared to when only one modality is hateful (and the other has been replaced by a benign variant). A high $\text{MIDAS}_{\text{text}}$ and low $\text{MIDAS}_{\text{cross}}$ indicate that the model is internally focusing its attention on text-only interactions (a *textual shortcut*). Conversely, a high $\text{MIDAS}_{\text{image}}$ and low $\text{MIDAS}_{\text{cross}}$ signal an *image shortcut*. A model that correctly balances both modalities will exhibit a relatively large $\text{MIDAS}_{\text{cross}}$, reflecting true cross-modal synergy.

In practice, we compute these three scalar MIDAS values for each hateful-benign triplet in our evaluation subset. Formal relationship between MITE and MIDAS is discussed in Appendix A.1.

3.5.5 Bias-Aware Calibration

We propose a lightweight calibration method that mitigates modality bias by incorporating attention-based attribution features. This procedure complements standard predictions with interaction-type-specific attribution scores, enhancing robustness against unimodal shortcuts. Our calibration pipeline

builds on attention attribution scores extracted from the final layer of each VLM. For each hateful-vs-benign triplet (T, I) , we compute the following:

- The predicted probability $\theta_{T,I}$,
- Attention attributions $\text{attr}_{i,j}$ for all token-region pairs (i, j) ,
- Aggregated scores across three interaction types:

$$\begin{aligned} - \text{attr}_{\text{text}} &= \sum_{(i,j) \in \mathcal{P}_{\text{text}}} \text{attr}_{i,j}, \\ - \text{attr}_{\text{image}} &= \sum_{(i,j) \in \mathcal{P}_{\text{image}}} \text{attr}_{i,j}, \\ - \text{attr}_{\text{cross}} &= \sum_{(i,j) \in \mathcal{P}_{\text{cross}}} \text{attr}_{i,j}. \end{aligned}$$

We then train four LightGBM classifiers, each using a different feature set for comparative evaluation:

1. **Raw-Only Baseline:** uses only $\theta_{T,I}$.
2. **Single Attr Baseline** [3]: uses $\theta_{T,I}$ and aggregate attention $\sum A$.
3. **Two Score Baseline** [4]: uses $\theta_{T,I}$, $\text{attr}_{\text{text}} + \text{attr}_{\text{cross}}$, and $\text{attr}_{\text{image}} + \text{attr}_{\text{cross}}$.
4. **Three Interaction Types (Proposed):** uses $\theta_{T,I}$ with the full interaction set $\{\text{attr}_{\text{text}}, \text{attr}_{\text{image}}, \text{attr}_{\text{cross}}\}$.

Each classifier is evaluated on its ability to improve prediction quality in the presence of modality shortcuts. Among these, only the proposed and attribution-informed variants perform bias-aware calibration by leveraging internal attention signals; the raw-only baseline lacks any modality-awareness and thus cannot correct for modality imbalance.

3.6 Experimental Settings

This section details the data preparation, model configurations, and evaluation protocol used to (i) compute MITE and MIDAS for each VLM, and (ii) train and assess the bias-aware calibration module. In summary, we:

1. For fine-tuned Oscar, UNITER, and VisualBERT in Vilio’s repository¹, we compute MITE and MIDAS on the *dev-seen* set (162 hateful memes, each paired with image- and text-benign confounders).
2. We analyze model behavior by comparing *Text-to-Text* and *Image-to-Image* settings separately, measuring how hate predictions and attention distributions change under each type of modality manipulation.
3. Finally, we train a *bias-aware LightGBM* re-ranker using attribution features and compare classification accuracy against the raw model baseline.

¹<https://github.com/Muennighoff/vilio>

3.6.1 Dataset and Confounder Extraction

We conduct all experiments on the HMC dataset, which contains 10 K crowd-sourced image-text pairs. Among these, there are 162 *dev-seen* examples labeled as *hateful*. Each hateful meme $(T_1^{\text{org}}, I_1^{\text{org}})$ comes with exactly one *provided* unimodal benign confounder: either image-benign $(T_1^{\text{org}}, I_0^{\text{org}})$ or text-benign $(T_0^{\text{org}}, I_1^{\text{org}})$.

To enable consistent computation of MITE and MIDAS across all examples, we construct two complete evaluation settings per hateful meme by sampling additional benign inputs when needed:

T2T (Text-to-Text) Task: Compare original hateful sample $(T_1^{\text{org}}, I_1^{\text{org}})$, $(T_1^{\text{org}}, I_0^{\text{org}})$, a randomly picked text benign $(\tilde{T}_0, I_1^{\text{org}})$.

I2I (Image-to-Image) Task: Compare original hateful sample $(T_1^{\text{org}}, I_1^{\text{org}})$, $(T_0^{\text{org}}, \tilde{I}_0)$, a randomly picked text benign $(T_0^{\text{org}}, I_1^{\text{org}})$.

To reduce sampling bias and ensure each hateful meme has both confounder types, we fill in any missing modality with three distinct *picked* benign samples from the corresponding unimodal pool:

- For 84 memes with provided text-benign confounders, we sample three benign images $\tilde{I}_0^{(1)}, \tilde{I}_0^{(2)}, \tilde{I}_0^{(3)}$ to form:

$$(T_1^{\text{org}}, \tilde{I}_0^{(1)}), \quad (T_1^{\text{org}}, \tilde{I}_0^{(2)}), \quad (T_1^{\text{org}}, \tilde{I}_0^{(3)}).$$

- For 78 memes with provided image-benign confounders, we sample three benign texts $\tilde{T}_0^{(1)}, \tilde{T}_0^{(2)}, \tilde{T}_0^{(3)}$ to form:

$$(\tilde{T}_0^{(1)}, I_1^{\text{org}}), \quad (\tilde{T}_0^{(2)}, I_1^{\text{org}}), \quad (\tilde{T}_0^{(3)}, I_1^{\text{org}}).$$

Sample Category	# of Samples
Hateful $(T_1^{\text{org}}, I_1^{\text{org}})$	162
Provided Image-Benign $(T_1^{\text{org}}, I_0^{\text{org}})$	78
Provided Text-Benign $(T_0^{\text{org}}, I_1^{\text{org}})$	84
Picked Image-Benign $(T_1^{\text{org}}, \tilde{I}_0)$	234
Picked Text-Benign $(\tilde{T}_0, I_1^{\text{org}})$	252

Table 3.2: Breakdown of all hateful and benign pairs. Each hateful meme is paired with one provided benign example and three picked confounders to ensure both modalities are covered.

In total, we obtain $162 + (78 + 84) + (234 + 252) = 810$ distinct pairs.

Each hateful meme thus yields two complete triplets:

$$\begin{aligned} \text{T2T: } & (T_1^{\text{org}}, I_1^{\text{org}}), \quad (T_1^{\text{org}}, I_0^{\text{org}}), \quad (\tilde{T}_0, I_1^{\text{org}}), \\ \text{I2I: } & (T_1^{\text{org}}, I_1^{\text{org}}), \quad (T_1^{\text{org}}, \tilde{I}_0), \quad (T_0^{\text{org}}, I_1^{\text{org}}). \end{aligned}$$

We compute MITE as:

$$\text{MITE}_{\text{T2T}} = \theta_{T_1^{\text{org}}, I_1^{\text{org}}} - \left(\theta_{T_1^{\text{org}}, \tilde{I}_0} + \theta_{T_0^{\text{org}}, I_1^{\text{org}}} \right),$$

and analogously for MITE_{I2I} . Larger discrepancies between these values indicate modality-specific shortcut behavior. To protect against confounder noise, each MITE or MIDAS computation is averaged over three independently sampled benign inputs per missing modality.

3.6.2 Models

We leverage the Vilio codebase, which provides author implementations of three canonical VLM fusion architectures—Oscar, UNITER, and VisualBERT—each fine-tuned on the HMC training set. Below is a summary of their encoder backbones and pretraining objectives (Table 3.3). In all cases, they use the basic BERT encoder ($\text{BERT}_{\text{base}}$) to embed text tokens and a Faster-R-CNN visual-feature extractor. For each architecture, we actually

Model	Encoder	Pretraining Tasks
Oscar	BERT (base)	(1) Object-Tag Detection (2) Image-Text Contrastive Learning
UNITER	BERT (base)	(1) Masked Language Modeling (2) Masked Image Modeling (3) Image-Text Matching (4) Word-Region Alignment via Optimal Transport
VisualBERT	BERT (base)	(1) Masked Language Modeling (2) Image Captioning

Table 3.3: Summary of the three models used in our experiments.

evaluate *three* submodels (each differing by either random seed, the number of visual features, or pretraining corpus variant—see the author’s repository²

²https://github.com/Muennighoff/vilio/blob/master/SCORE_REPRO.md

for more detail), but empirical trends for visualization (MITE gaps and MIDAS patterns) are qualitatively consistent across all submodels. To keep the presentation concise, we report results from one representative submodel per architecture. For the calibration experiment, we report the results from all submodels.

3.6.3 Bias-Aware Calibration Set-up

To evaluate calibration efficacy, we construct a dataset of $162 \times 3 = 486$ samples, where each hateful meme is paired with one image-benign and one text-benign variant. For every triplet, we extract the full feature set and assign a binary label (hateful vs. benign). The LightGBM classifiers are trained with the following configuration:

- 1000 boosting rounds (trees),
- Maximum depth of 6,
- 80% training and 20% validation split,
- Five fixed random seeds: {42, 123, 999, 1221, 1987}.

We report macro-AUC averaged across five random splits.

3.6.4 Implementation Details

All experiments were conducted using PyTorch 1.12, Transformers 4.21, and LightGBM 3.3, running on NVIDIA A100 80GB GPUs.

3.7 Results & Discussion

We present three experiments to evaluate modality biases in SotA hateful meme detectors and explore bias mitigation via attr.

3.7.1 Experiment I: Modality Bias in Prediction (MITE)

We compute MITE in both *Text-to-Text* (T2T) and *Image-to-Image* (I2I) settings for each model on the HMC dev-seen split. These values reflect how much the model’s hate prediction relies on joint text-image synergy, as opposed to unimodal cues. Figure 3.2 displays the average MITE scores for Oscar (O), UNITER (U), and VisualBERT (V), shown separately for the T2T (cyan) and I2I (magenta) settings. VisualBERT shows a substantially higher score in T2T than in I2I, indicating stronger reliance on the text modality. Oscar exhibits a more balanced pattern, with moderate preference toward

text. UNITER shows nearly equal scores across both settings, suggesting that it relies on both modalities more evenly. These trends suggest that

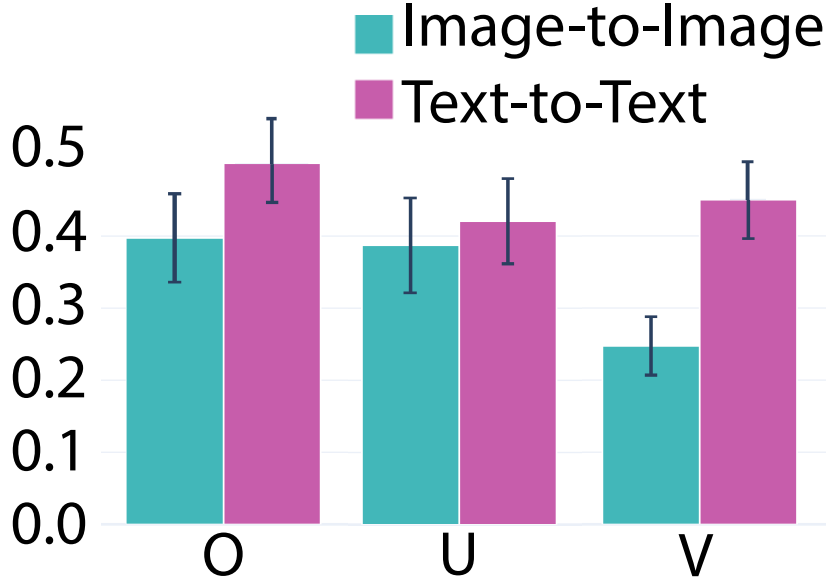


Figure 3.2: MITE across Oscar (O: left), UNITER (U: middle), and VisualBERT (V: right), shown separately for *Image-to-Image* (magenta) and *Text-to-Text* (cyan) settings. Larger MITE values indicate greater reliance on joint text-image signals. A notable gap between the two bars suggests asymmetric reliance on one modality.

pretraining and architecture choices influence how different models integrate multimodal information, and shortcut on text-only signals in some cases.

3.7.2 Experiment II: Modality Bias in Attention: MIDAS

3.7.2.1 Global Analysis

To inspect internal attention level biases, we calculate MIDAS values for each model under I2I and T2T settings. Figure 3.3 presents the average $\text{MIDAS}_{\text{text}}$, $\text{MIDAS}_{\text{image}}$, and $\text{MIDAS}_{\text{cross}}$ scores for Oscar, UNITER, VisualBERT, and an ablated VisualBERT variant pretrained solely on text. The result shows that VisualBERT’s attention heavily favors text interactions (Figure 3.3, third row). Oscar displays a moderate $\text{MIDAS}_{\text{text}}$ $\text{MIDAS}_{\text{cross}}$ difference, reflecting an intermediate textual shortcut (first row). UNITER

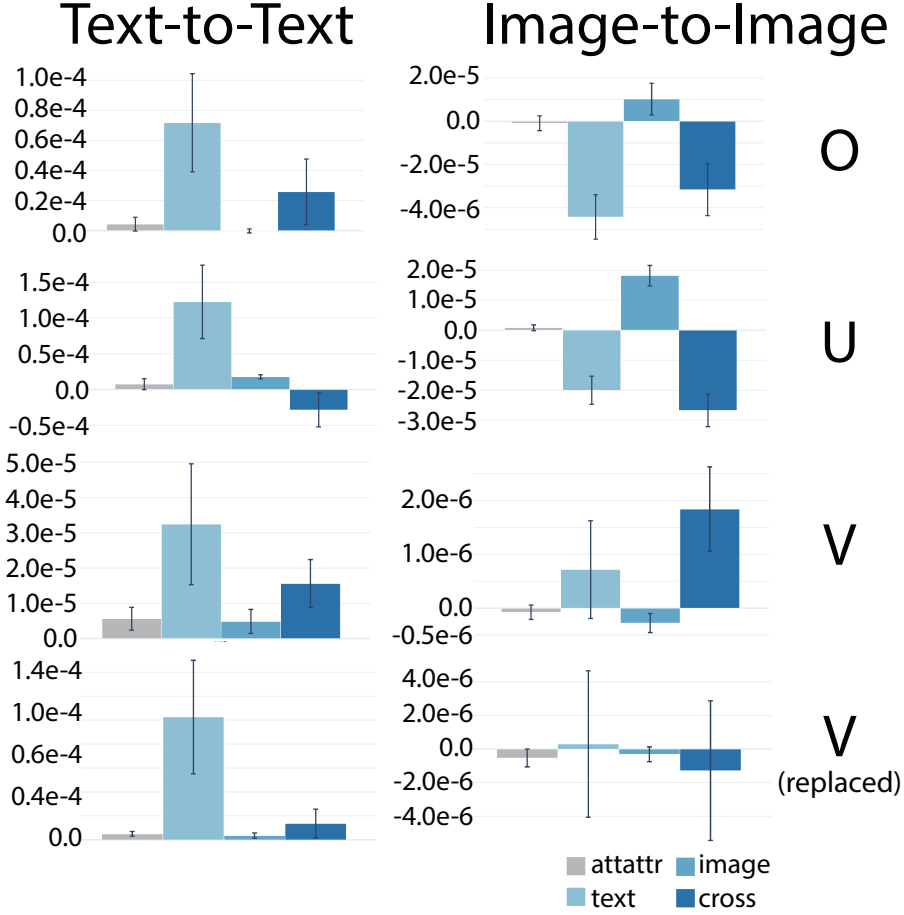


Figure 3.3: Global MIDAS for Text-to-Text (left) and Image-to-Image (right) samples, for Oscar (top), UNITER (second row), VisualBERT (third row), and VisualBERT (text-only pretrained) (bottom). Bars in each graph show attr summed over all the elements, $\text{MIDAS}_{\text{text}}$, $\text{MIDAS}_{\text{cross}}$, and $\text{MIDAS}_{\text{image}}$ (from left to right). Consistently with MITE result, VisualBERT shows the strongest reliance on $\text{MIDAS}_{\text{text}}$ for both tasks, while other models relies more on $\text{MIDAS}_{\text{image}}$ and $\text{MIDAS}_{\text{cross}}$ for the Image-to-Image task.

exhibits the smallest such gap, suggesting its attention is more evenly distributed across modalities (second row). Thus, MIDAS confirms the ordering $\text{VisualBERT} > \text{Oscar} > \text{UNITER}$ in terms of internal reliance on textual interactions. Additionally, the text-only pretrained VisualBERT variant (fourth row) further amplifies this bias, confirming that pretraining data significantly influences modality interactions.

3.7.2.2 Local Analysis

We further examine local explanations for representative hateful–benign meme pairs to understand fine-grained attribution differences qualitatively. Figure 3.4 illustrates a sample where UNITER highlights semantically relevant regions (e.g. the woman vs. cargo in the image) and tokens (e.g. *dishwasher* vs. *driving*). The top-scored image and text features differ



Figure 3.4: Local MIDAS explanations for a representative hateful (row 1), text-benign (row 2), and image-benign (row 3) sample using UNITER. Top image regions and text tokens receive the highest attribution under each intervention. UNITER attends to "dishwasher" in the text and a woman in the image, suggesting that this model successfully capture the hateful analogy.

when intervening on each modality, demonstrating UNITER's ability to

capture multimodal contrasts. Models exhibiting stronger textual biases, like VisualBERT, underweight critical visual cues and fail to highlight salient image regions (Appendix A.2).

3.7.3 Experiment III: Calibration with Attention Attribution

To evaluate whether attr_{it} can mitigate modality bias and boost classification performance, we train LightGBM classifiers on four feature sets (§3.6.3). Table 3.4 summarizes the results, showing AUROC scores for each model and feature set. The result shows that VisualBERT achieves the largest

Table 3.4: Calibration results. The first column lists the model, followed by the submodel identifier. The next three columns show AUROC scores for the raw prediction only, single attention attribution [3], and two score baseline [4]. The last column shows the performance of our proposed three-interaction-type method. Performance is evaluated by AUROC, shown in percentage for brevity. For Oscar and VisualBERT, our three-interaction-type method outperforms single or dual attention split baselines with an exception of O36.

Model	Sub-model	Pred-Only	Single Attr.	Two Score	attr_{it}
O	O36	80.0%	81.3%	79.1%	81.2%
	O50	83.7%	86.3%	84.4%	86.4%
	OV50	83.8%	84.3%	82.2%	85.1%
U	U36	87.3%	83.2%	88.3%	85.9%
	U50	83.6%	82.4%	84.8%	82.0%
	U72	82.3%	79.5%	80.9%	79.3%
V	V135	79.1%	80.9%	79.3%	81.8%
	V45	79.1%	80.9%	79.3%	81.8%
	V90	79.1%	80.9%	79.3%	81.8%

boost ($\simeq +2.7$ points), reflecting how much its raw predictions suffered from text-only shortcuts. Oscar sees a substantial gain ($\simeq +2.7$ points on O50), confirming its intermediate bias. UNITER shows negligible or even negative improvements, consistent with its minimal textual bias in MITE and MIDAS. Additionally, our three-interaction-type approach consistently outperforms previously proposed metrics (aggregated attr [3] and text-image bimodal scoring [4]) for these biased models. These results indicate that

attr_{it} effectively corrects modality biases by injecting cross-modal interaction signals in case a VLM’s decisions are biased towards textual modality.

3.7.4 Discussion

Our two-fold causality-interpretability framework combining MITE with MIDAS reveals how hateful meme detectors shortcut on single modalities at both the output and attention levels.

First, by measuring changes in hate-probability when swapping text or image inputs, MITE quantifies the model’s reliance on joint versus unimodal cues. As shown in Figure 3.2, MITE gap exhibits a clear gradient: largest gap in VisualBERT, moderate in Oscar, no significance in UNITER, quantifying the reliance on text when the image changes.

Next, MIDAS decomposes last-layer attention into within-text, within-image, and cross-modal contributions. Global analysis (Figure 3.3) mirror the MITE’s pattern: the within-text minus cross-modal attribution is highest for VisualBERT, moderate for Oscar, and lowest for UNITER. Additionally, inspecting representative hateful–benign triplets, local MIDAS explanations reveal that UNITER highlights semantically relevant regions and tokens under each intervention, whereas models with stronger unimodal biases underweight complementary modality cues (Figure 3.4 and Appendix A.2). Finally, the calibration gain obtained by injecting the three attr_{it} features into a lightweight LightGBM re-ranker (Table 3.4) peak for VisualBERT, follow for Oscar, and are minimal for UNITER, showing that the calibration is most needed (and most effective) where text bias was strongest.

Together, these results highlight the value of diagnosing both output- and attention-level shortcuts when evaluating multimodal HM detectors, rather than relying solely on accuracy metrics. However, our analysis is confined to the controlled HMC confounders and three fusion architectures; real-world memes and other VLMs may present more complex bias patterns. Future work could extend MITE and MIDAS to larger and more diverse datasets, explore automated confounder generation, and investigate how mechanistic interventions in intermediate representations contribute to alleviating modality shortcuts.

Unlike [47], who show via Shapley-value analysis [75] that text alone often drives cross-domain generalization and that adding image captions can actually hurt VL models, MITE formally quantifies the *intersectional* effect of multimodal signals versus each modality in isolation, rather than relying on post hoc feature-importance methods. Similarly, whereas [4] split model attention into only text-only and image-only attributions—revealing slur grounding biases and visual keyword over-sensitivity—MIDAS adds a third

cross-modal category and then demonstrates how these fine-grained signals can be injected into a LightGBM re-ranker to mitigate unimodal shortcuts. Finally, [48] articulate a general causal interpretation of intersectionality in social theory. This work is the first to operationalize those ideas in an empirical, large-scale evaluation of VLMs on HMD.

As mentioned in §2.5.1, previous approaches for causally debiasing VLMs are limited to debiasing predictions, not attention or other internal bias indicators. To further improve our debiasing methodologies, however, our method can be combined with arbitrary methods from this approach.

3.7.5 Limitations

This study’s primary limitation concerns the unverified generalizability of its findings. Hateful memes represent an evolving area of concern that necessitates extensive, openly accessible datasets for comprehensive analysis and validation. Although this chapter provides insights into bias detection and mitigation in HMD, the framework should be tested on more datasets and problem settings, such as multi-class meme detection [76].

A further constraint is the linguistic homogeneity of the dataset employed, with the HMC dataset comprising exclusively English-language textual content. This presents a critical limitation in the context of the global escalation of extremism, where hate speech proliferates across linguistic boundaries. The detection of multilingual hate speech thus emerges as a crucial area for future research, necessitating methodologies capable of navigating language-specific nuances and cultural contexts [77, 78].

Additionally, the field of hate speech detection faces resource limitations, notably in the size and diversity of available datasets. Hateful speech datasets are generally small, restricting the depth and breadth of training data for machine learning models. We believe future studies could utilize LLMs and VLMs as dataset curators (see a prompt engineering solution in Chapter 4). Although MIDAS provides an interpretable high-level summary of the attention attribution to the predicted hatefulness, it shows limited performance gain in debiasing the baseline classifiers. A potential source of this limitation could be alleviated by combining lower-level approaches, such as manipulating the raw attention matrix via MI methodologies.

From a causal perspective, we should keep in mind that the causality presupposed by MITE and MIDAS is *unconditioned*: although we believe that the central idea (a text and an image are combined with malicious intention to form the indirect effect, generating a hateful label) is logically sound, it is *challenging to validate this idea itself*. In fact, the concept of hatefulness in a meme is rather complicated; for example, the original HMC

dataset contains unimodal hateful meme³ [41], indicating the presence of another path from text or image. More fine-grained analysis on the type of meme should be a part of the future work.

3.8 Conclusion

In this chapter, we presented a unified, causality-and-interpretability-driven framework for diagnosing and mitigating modality bias in hateful meme detectors. We first introduced *MITE*, a causal metric to measure how models leverage joint text-image cues, and then developed *MIDAS*, an attention-based metric to pinpoint where in the network these shortcuts occur. Applied to three SotA VLMs, our analyses uncovered significant textual biases, especially in VisualBERT, and demonstrated that injecting MIDAS features into a bias-aware LightGBM re-ranker consistently improves detection performance for modality-biased models. Together, these findings advocate for a shift from purely accuracy-centered evaluation to causal analyses in multimodal modeling. Future work will extend this pipeline to additional datasets, explore other forms of presentation bias, and investigate automated confounder generation via LLMs.

³For example, when the text abusing people with a specific race is embedded in the normal image of that race, this meme is unimodal (the message is solely encoded in single modality, the text.)

Chapter 4

Formatting Bias within a VLM for VQA

4.1 Terminology

Table 4.1: Definitions of key terms.

Term	Definition
Representational Shift	Changes in the model’s internal representations (hidden states) caused by In-Context Learning (ICL) examples, reflecting how examples reshape zero-shot processing.
Representational Shift Theory (RST)	Formal framework of representational shifts induced by ICL examples.
In-Distribution (ID)	Data formatted like training (single-image, single-turn). In this chapter, ID is identical to Zero-Shot Learning (ZSL), a setting where a model is required to solve a task without In-Context Learning (ICL) examples.
Out-of-Distribution (OoD)	New formats (multi-image, multi-turn).
OoD ICL	Providing in-context examples in formats not seen during training to test semantic generalization.
Multi-image Multi-turn (M^2 OoD)	The inputs presented in multi-image multi-turn fashion whereas the model is trained on single-image or single-turn datasets.

continued on next page

Table 4.1 – *continued from previous page*

Term	Definition
Single-image Multi-turn OoD (SM OoD)	The inputs presented in single-image multi-turn fashion whereas the model is trained on single-image or single-turn datasets.
Counterfactual Prompting (CFP)	A method of generating in-context examples by prompting the model to produce alternative captions or responses, used to mitigate bias.

4.2 Chapter Summary

In this chapter, we address formatting bias in generative VLMs. First, we motivate the importance of Out-of-Distribution (OoD) generalization (§4.3). In OoD inputs, the same semantic content appears under new formats. To generalize, a VLM must mitigate the *formatting bias* introduced by its training format. We focus on how SotA VLMs tackle this OoD generalization when an ICL example is presented in an OoD fashion (OoD ICL).

Next, we review relevant literature (§4.4) on multimodal prompt engineering, bias measurement in VLMs, and theoretical accounts of ICL, identifying a gap in principled frameworks that explain how examples reshape internal representations to drive robust OoD behavior.

To address this, we introduce RST (§4.5.3). We formalize the representational shift vector between zero-shot and in-context representations, and propose two hypotheses: *Hypothesis I*, that semantically rich multimodal OoD examples enhance performance when zero-shot semantics are weak; and *Hypothesis II*, that text-only OoD examples recover performance in text-dominated tasks like HMD.

We then detail our experimental settings (§4.6): evaluating on six VQA benchmarks [39, 79–83] with two Large Language and Vision Assistant (LLaVA) [2] variants; describing prompt templates including counterfactual injections; and outlining how we extract representations and compute shifts. In the Results & Discussion (§4.7), we present two experiments confirming our hypotheses. In *Experiment I* (§4.7.1), semantically rich multimodal examples yield significant performance gains, especially on tasks with low zero-shot accuracy. In *Experiment II* (§4.7.2), textual CounterFactual Prompting (CFP) Framework recovers F1-score losses and mitigates label bias in the HMC dataset. We synthesize these findings to illustrate how RST guides effective prompt design for mitigating formatting biases in OoD ICL.

Finally, we summarize our contributions (§4.8): a theory-driven framework

for interpretability via representational shifts, empirical validation of our hypotheses across VQA and HMD tasks, and practical guidelines for ICL example selection. We conclude by outlining future directions: automating hypothesis curation, RepE for driving the shifts in latent space, and LLM applications.

4.3 Introduction

Large VLMs have made remarkable strides on many multimodal tasks, yet their robustness under distributional shifts remains a pressing concern. A key challenge is *formatting bias mitigation*: extracting true semantic content independently of spurious statistical regularities in formatting or prompt structure. When prompts or inputs shift beyond the training distribution—whether through novel combinations of images and text, or through subtle alterations in prompt phrasing—models that have not obtained such a mitigation ability can produce unpredictable or erroneous outputs. Ensuring reliable OoD generalization is therefore critical for deploying generative VLMs in real-world applications where inputs rarely mirror the training data. Existing work has explored various aspects of this problem. Multimodal prompt engineering has demonstrated that example selection can steer model behavior, and empirical studies have quantified biases arising from various shortcuts in VLMs. Theoretical accounts of ICL have shed light on how demonstration examples influence internal representations. However, these efforts remain fragmented: there is no unified, mechanistic framework that explains *how* and *why* particular examples reshape a model’s latent space to promote robust OoD performance, nor clear guidelines for constructing demonstrations that systematically induce desirable shifts for mitigating formatting biases.

In this chapter, we introduce *RST*, a principled framework for interpreting and guiding example-induced changes in VLMs. We formalize the *representational shift vector* which measures the difference between the in-context representation produced by an ICL example and the zero-shot representation of a test input. We posit two complementary hypotheses:

1. **Hypothesis I.** When zero-shot semantics are weak, semantically rich, multi-image multi-turn demonstrations will yield substantial performance gains for a VLM solely trained with single-image single-turn datasets (Multi-image Multi-turn OoD (M^2 OoD)).
2. **Hypothesis II.** When textual content predominates, carefully crafted text-only demonstrations can recover performance losses and mitigate label bias for a VLM solely trained with single-image single-turn

datasets (Single-image Multi-turn OoD (SM OoD)).

Our evaluation across six VQA benchmarks and the HMC dataset with two LLaVA variants confirms these hypotheses: in Experiment I, multimodal OoD examples deliver significant accuracy improvements, particularly on the tasks with low Zero-Shot Learning (ZSL) performance; in Experiment II, text-only CFP recovers F1 losses and effectively reduces label bias on HMC data.

Our contribution is summarized as:

- **Testable ICL Theory on Formatting Bias Mitigation.** We propose RST, a theory-driven framework that explains how in-context examples induce targeted changes in a VLM’s latent space. We further formalize the representational shift vector and derive two core hypotheses (Hypothesis I about M^2 OoD and Hypothesis II about SM OoD) for guiding example selection.
- **Empirical Validation in VQA.** We empirically validate the Hypothesis I on six standard VQA benchmarks, demonstrating significant accuracy improvements from semantically rich multi-image multi-turn demonstrations even though multi-image format is not presented in the training (M^2 OoD).
- **Empirical Validation in HMD.** We empirically validate the Hypothesis II on the HMC dataset, showing that single-image multi-turn text-only counterfactual prompting recovers performance and reduces label bias even though the models are not trained in ICL fashion (SM OoD).
- **Practical ICL Guidelines in OoD settings.** We present practical guidelines for constructing effective demonstrations in generative VLMs facing OoD problems, grounded in measured representational shifts.

4.4 Related Work

In this section, we review three strands of research central to our study: (1) the pursuit of separating how semantic information disentangled from formatting bias in model representations; (2) the mechanisms and theories underpinning ICL; and (3) prior work on OoD generalization in VL tasks.

4.4.1 Towards Formatting Bias Mitigation

Extracting meaningful content independently of superficial statistical patterns, or formatting bias, is the central question of interpretability studies.

Prior research has typically isolated individual semantic factors, such as color perception stripped of spectral information [84], geolocation under scaling laws [85–87], or spatial world model [88], but lacks a unified framework spanning multiple aspects of the semantics. Our work builds on these foundational efforts of semantic information extraction to establish a theory-driven approach to *mitigate* formatting bias for maximizing the VLM’s understanding of the tasks at hand.

4.4.2 In-Context Learning (ICL)

Since its emergence in GPT-3 [12] as a gradient-free adaptation mechanism, ICL has attracted extensive study. Towards unleashing the full capabilities of LLMs and VLMs, efforts on optimizing the ICL strategies divide into three categories: prompt engineering (e.g. chain-of-thought prompts) [89]; example selection via retrieval methods like BM25 [90] or CLIP embedding similarity [91]; and curated demonstration generation. Notably, CFP grounded ICL example selection in the data generation process [92].

Concurrently, studies on interpretability have probed ICL dynamics through specific algorithms, such as Bayesian inference [93], contrastive learning [94], multi-state RNN [95], and gradient descent [96, 97], among many others [98–100]. These studies covered extensive theoretical aspects, including the common finding of *meta-gradient*; LLMs could learn how to optimize its own representation. However, how each theory contributes to separate semantics from formatting bias is unclear.

As more empirical approaches for understanding how the ICL works, MI works discovered key circuits and mechanisms behind ICL, such as induction heads [101, 102], function vectors [103], and parallel structure [104]. Instead of focusing on the detailed mechanisms, our study provides a macroscopic analysis of the entire latent space. A potential mechanism behind our findings is discussed in the later section (§4.7.3).

Taking the mixed strategy of theoretical and empirical approaches, this paper leverages multi-image, multi-turn contexts to examine how varying semantics-format mixtures in demonstrations affect ICL efficacy, positioning our study as the first to link ICL theory directly with ICL performance.

4.4.3 Out-of-Distribution (OoD) Generalization

OoD generalization examines model robustness when test inputs diverge from training distributions [105, 106]. Traditional approaches in VL have focused on domain adaptation [107] and label-shift scenarios [108, 109]; however, these do not explicitly distinguish semantics from formatting bias; i.e., how

to define the distributional difference among multiple datasets or multiple labels is opaque. Recent work has begun to address OoD in multimodal conversations in single-image and multi-turn settings, and has proposed information-theoretic and label-shift benchmarks. For example, [110, 111] proposed solutions for detecting OoD in a multimodal conversation. [112] introduced the information-theoretic approach for multi-turn conversation intention detection. [113] proposed two novel OoD categories, the multi-label OoD and the label shift under the specific context. [114] benchmark OoD detection in VQA and show that attention-based scoring, specifically average maximum attention probability, effectively flags ungrounded or misaligned image-question pairs by capturing the model’s reduced attention alignment in OoD cases. We extend this line of research by evaluating and improving OoD generalization under multi-image and multi-turn prompts, offering new insights into how semantic information drives robust generalization.

4.5 Methodology

4.5.1 Proposed Methodology Overview

Our approach studies how ICL examples can induce targeted shifts in a VLM’s latent space to promote OoD generalization. First, as preliminaries to our methodology (§4.5.2), we introduce key concepts and two illustrative examples (*Banana Mustache* and *Tomato Nose*). Second, we formalize the RST, which decomposes these shifts into semantic and superficial format components (§4.5.3; see Figure 4.1 for an overview). Finally, we propose two hypotheses on OoD generalization, which we test in the experiments (§4.5.4) (Figure 4.2).

4.5.2 Preliminaries

4.5.2.1 Meta-Gradient

Central to traditional machine learning is gradient descent, where the learning objective induces a change ΔH in the hidden representation H . Recent interpretability studies suggest that LLMs perform an analogous, *implicit* form of gradient descent during ICL: they use their own attention weights W to compute a *meta-gradient* ΔW , which yields an updated representation

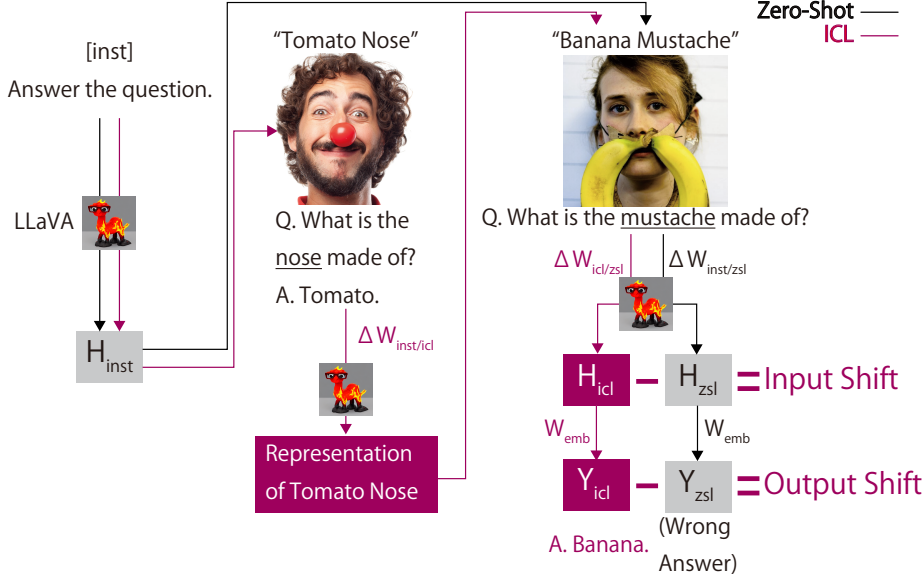


Figure 4.1: Schematic Overview of Representational Shift Theory (RST) with *Banana Mustache* test input and *Tomato Nose* In-Context Learning (ICL) example. Both in a zero-shot setting (ZSL; black line) and in ICL (red line), an instruction (question) is first provided with LLaVA to compose the representation H_{inst} . Next, in a zero-shot setting, the test input forms the meta-gradient $\Delta W_{\text{inst/ZSL}}$, resulting in the test-input’s representation H_{ZSL} . Finally, the unembedding weight W_{emb} converts H_{ZSL} to the output Y_{ZSL} . In contrast, in ICL, an example is inserted between the instruction and the test input to contextualize the semantic components (analogy of fruits and facial parts) to form the meta-gradient $\Delta W_{\text{inst/icl}}$ and $\Delta W_{\text{icl/zsl}}$, and then the ICL example’s representation H_{ICL} and the output Y_{ICL} . We argue that the difference of test-input representation $H_{\text{ICL}} - H_{\text{ZSL}}$ (input shift) and that of output representation $Y_{\text{ICL}} - Y_{\text{ZSL}}$ (output shift) are the core of ICL.

H' when applied to H .

$$\begin{aligned}
 \underbrace{H_{\text{ZSL}}}_{\text{zero-shot}} &= (W - \Delta W_{\text{inst/zsl}}) H_{\text{inst}} \\
 \underbrace{H_{\text{ICL}}}_{\text{in-context}} &= (W - (\Delta W_{\text{inst/icl}} + \Delta W_{\text{icl/zsl}})) H_{\text{inst}}
 \end{aligned} \tag{4.1}$$

Here, H_{inst} is the model’s hidden representation of the task instruction inst . $\Delta W_{\text{inst/zsl}}$ is the meta-gradient induced by the instruction alone in ZSL. $\Delta W_{\text{inst/icl}}$ and $\Delta W_{\text{icl/zsl}}$ are the successive meta-gradients induced by the instruction→example and example→test mappings in ICL.

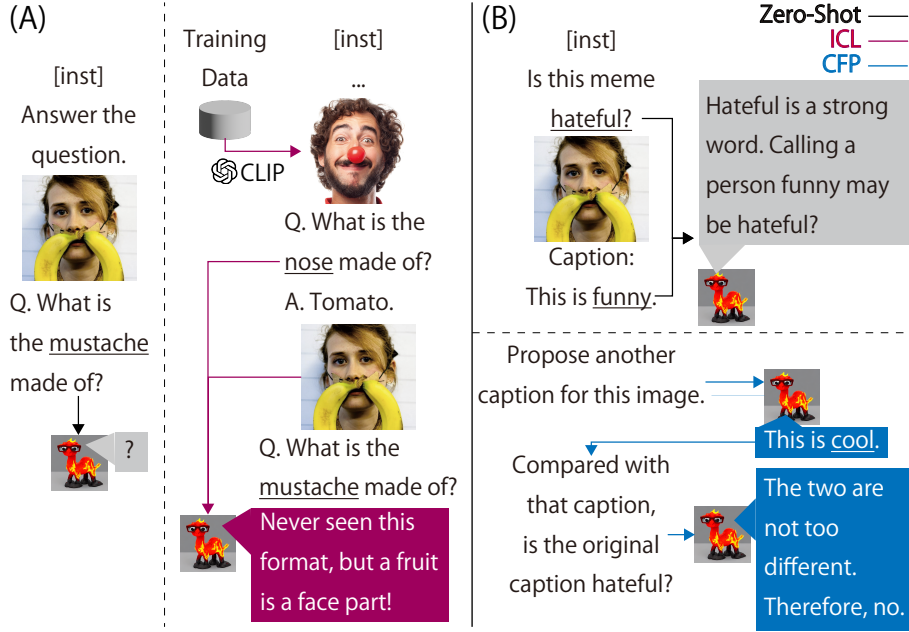


Figure 4.2: Schematic Drawings of Hypotheses I/II and Experiments I/II. (A) Hypothesis I and Experiment I. We hypothesize that semantically rich ICL example given in an unfamiliar format contributes to the performance when the zero-shot performance is poor (left). To validate this hypothesis, we tested LLaVA’s Out-of-Distribution (OoD) ICL performance by adding an ICL example most similar to the test input based on CLIP embedding (semantically rich) to the LLaVA variants *not* trained with multi-image datasets (right). When the model successfully mitigate formatting bias, it extracts the *semantic* analogy despite the unseen format (large and explicit *formatting bias*). (B) Hypothesis II and Experiment II. When LLaVA is textually biased towards the hateful label (top), we hypothesize that enhancing text-to-text interaction facilitates the unbiased decision (bottom), and test this hypothesis with counterfactual prompting (instructing the model to propose a caption to compare with the original caption).

4.5.2.2 Unembedding

Once a hidden state H is computed (whether in ZSL or ICL setting) the model’s outputs Y are obtained by a linear projection (*unembedding*) via a weight matrix W_{emb} :

$$Y = W_{\text{emb}} H \quad (4.2)$$

This operation maps the high-dimensional latent representation back into the model’s output vocabulary or label space. With these definitions in

place, we can precisely quantify how ICL examples induce shifts in both the input representation ($H_{\text{ICL}} - H_{\text{ZSL}}$) and the generative outputs ($Y_{\text{ICL}} - Y_{\text{ZSL}}$), forming the foundation of RST.

4.5.2.3 Illustrative Examples

Before moving on to the mathematical details, we present two concrete examples which are used in the later sections for illustrative purposes.

1. Banana Mustache example:

(a) **Experiment I (VQA):**

Image: a face with a banana arranged as a mustache.

Question: “What is the mustache made of?”

Here, the task is to identify the object in the designated region.

(b) **Experiment II (HMD):**

Image: (same as Experiment I)

Question: “Is this image hateful or benign?”

In this binary moderation setup, the task is HMD.

2. Tomato Nose example:

(a) **Experiment I (VQA):**

Image: a face with a tomato arranged as a nose.

Question: “What is the nose made of?”

Here, the task is to identify the object in the designated region.

Intuitively, the *Tomato Nose* example is semantically similar to the *Banana Mustache* example, both involving a face and an object in the designated region.

Note that the *Tomato Nose* analogy is only employed to illustrate how an ICL demonstration induces a representational shift, and it is *not* in the real dataset. *Banana Mustache* is from the VQA dataset [39].

4.5.3 Representational Shift Theory (RST)

In short, RST states that an ICL example triggers the cascadic shifts of input representation and output. The core component of RST is the *representational shift vector*: the difference between in-context and zero-shot representations, enabling us to analyze how ICL examples induce targeted changes in a VLM’s latent space. Figure 4.1 illustrates the core components of RST. In the following sections, we will define the representational shift vector and the cascadic shifts triggered by an ICL example, and then introduce two hypotheses on OoD generalization.

4.5.3.1 Representational Shift Vector

First, we define the *representational shift vector* $\Delta R(E, x)$, which captures how an ICL example E reshapes the model’s representation of a test input x . This vector is defined as the difference between the in-context representation $R_{\text{ICL}}(E)$ and the zero-shot representation $R_{\text{ZSL}}(x)$.

$$\Delta R(E, x) = R_{\text{ICL}}(E) - R_{\text{ZSL}}(x) \quad (4.3)$$

Hereafter, we omit the input x and the example E when the context is clear, and simply denote the representational shift vector as ΔR . We decompose ΔR into a **semantic shift** (driven by content) and a **formatting shift** (driven by the number of images and turns)—the latter being the very artifact we seek to mitigate (Appendix B.1.2).

4.5.3.2 Cascadic Shifts

Specifically in the VQA task, R can be the hidden state H or the output Y of a VLM. The representational shift vector ΔR thus captures how the ICL example E modifies the model’s internal representation and final output for the test input x . We can express the representational shift vector in terms of the meta-gradient ΔW (Eq. 4.1) and unembedding weight W_{emb} (Eq. 4.2) as follows:

$$\begin{aligned} \text{Input Shift: } \Delta H &= H_{\text{ICL}} - H_{\text{ZSL}} \\ &= [W - (\Delta W_{\text{inst/icl}} + \Delta W_{\text{icl/zsl}})] H_{\text{inst}} - (W - \Delta W_{\text{inst/zsl}}) H_{\text{inst}} \\ &\propto -\Delta W_{\text{icl/zsl}} H_{\text{inst}} \\ \text{Output Shift: } \Delta Y &= Y_{\text{ICL}} - Y_{\text{ZSL}} \\ &\propto -W_{\text{emb}} \Delta W_{\text{icl/zsl}} H_{\text{inst}} \end{aligned} \quad (4.4)$$

which capture how semantic content in ICL examples reshapes both internal representations and final outcome. Here we assume that the meta-gradient triggered by the instruction is the same in both zero-shot and in-context settings (i.e. $\Delta W_{\text{inst/zsl}} \simeq \Delta W_{\text{inst/icl}}$). RST posits that these shifts are the key to understanding how ICL examples induce targeted changes in a VLM’s latent space, enabling robust OoD generalization.

4.5.4 Hypotheses on OoD Generalization

Building on RST, we posit two complementary hypotheses about how OoD ICL examples can steer VLM performance (Figure 4.2). These hypotheses are tested in §4.7.

4.5.4.1 Hypothesis I about Multi-Image Multi-Turn OoD

When a zero-shot (In-Distribution (ID)) input provides insufficient semantic cues, we hypothesize that inserting semantically rich ICL examples (resulting in M^2 OoD) will induce a representational shift that boosts performance (see Figure 4.2 (A)). For example, suppose a model trained solely with the single-image datasets (e.g. the images and captions of fruits or human faces) struggles with understanding the Banana Mustache example (Figure 4.2 (A), left) In this case, we provide a semantically rich OoD ICL example (Tomato Nose) to see if it can cause the meaningful representational shift, leading to better performance (Figure 4.2 (A), right).

4.5.4.2 Hypothesis II about Single-Image Multi-Turn OoD

In tasks prone to textual bias like HMD (Chapter 3), textual semantics dominate. In this case, we hypothesize that the *text-only* ICL examples (resulting in SM OoD) can mitigate textual bias and recover performance (see Figure 4.2 (B)). For example, a model instruction-tuned to flag aggressive messages with high recall may misclassify a neutral Banana Mustache meme due to spurious text cues (Figure 4.2 (B), top). By first generating a benign caption (e.g. “*This is cool*”) and then using that as an ICL example to compare against the original caption, we encourage text-to-text reasoning that reduces textual bias and improves HMD performance (Figure 4.2 (B), bottom).

4.6 Experimental Settings

4.6.1 Models

We evaluate two 13-billion-parameter LLaVA–Llama2 [2,115] and LLaVA–1.5 [116] chosen for strong language capabilities and *no* multi-image pretraining or multi-turn conversational instruction tuning.

4.6.2 Datasets

- **Experiment I (M^2 OoD):** We use six VQA benchmarks (VQA v2 [39], GQA [79], VizWiz [80], TextVQA [81], MMBench [82], MM-Vet [83]), spanning a wide range of zero-shot accuracies.
- **Experiment II (SM OoD):** We use HMD, selected for its text-dominant bias (Chapter 3).

```

System:
You are a helpful language and vision assistant.
User:
<image in dataset>
Give me one caption that fits with this image.
Assistant:
{generated caption}
User:
In comparison with that caption, is the following caption hateful or non-hateful?
Answer with a single word.
{caption in dataset}
Assistant:
{answer}

```

Figure 4.3: The representative counterfactual prompt. The system prompt is truncated for illustrative purposes. The curly brackets indicate the placeholders. This prompt instructs the model to generate a neutral caption that fits with the given image, and then to use that caption as a contrastive example for the hateful meme detection.

4.6.3 One-Shot Example Selection

- *Experiment I*: We retrieve the semantically closest multi-image example via CLIP embedding cosine similarity.
- *Experiment II*: We generate text-only demonstrations using CFP; compare against BM25 text retrieval (Search for the text most similar to the test input out of the training dataset, present the model with an image accompanied in the original dataset, utilizing the presence of text confounders observed in Chapter 3) as a baseline.

4.6.4 Prompt Templates

- *Experiment I*: We follow the zero-shot template in LLaVA’s repository¹. For M^2 OoD, we insert an example selected by the aforementioned criteria (§4.6.3). See §4.7.1.2 and Appendix B.4.2 for concrete examples.
- *Experiment II*: Figure 4.3 shows the template we used.

¹<https://github.com/haotian-liu/LLaVA>

4.6.5 Metrics and Representations

Following the official evaluation codes (e.g. VQAv2²), performance is measured via accuracy (VQA) in Experiment I. In Experiment II, we use F1 to balance the impact of each label since we hypothesize the bias towards a specific label (§4.5.4.2). Hidden states are extracted from the multimodal backbone; representational shifts are quantified by cosine similarity to link semantic content with performance changes.

In Experiment I, to disentangle dynamic semantic contributions from relatively static formatting biases in these shifts, we implemented a linear mixed-effect model [117]. We assumed that semantic effects vary across individual samples, whereas formatting biases (e.g. dataset-specific prompt formats or number of images) remain consistent within or across datasets. Accordingly, we built a linear mixed-effect model to predict the output shift vector ΔY (Eq. 4.4, bottom) from the input shift vector ΔH (Eq. 4.4, top). Specifically, we treated the input-shift vector as a random effect³, and added dummy variables for each of the six datasets and each LLaVA variant (LLaVA-Llama2 vs., LLaVA-1.5) as fixed effects⁴. Formal definitions and estimation details are provided in Appendix B.2.1, with further discussion on model choice in Appendix B.4.3. We use a random-effect-only model (i.e. without fixed effect) as a baseline.

In Experiment II, to directly assess how OoD ICL affects the separability of hateful versus benign representations, we compute a cosine similarity matrix over the input-shift vectors ΔH for all pairs of samples. Concretely, for each pair (i, j) with labels $y_i, y_j \in \{\text{hateful}, \text{benign}\}$, we calculate

$$\text{sim}(i, j) = \frac{\Delta H(i)^\top \Delta H(j)}{\|\Delta H(i)\| \|\Delta H(j)\|} \quad (4.5)$$

and organize these into a block matrix distinguishing intra-class (hateful-hateful, benign-benign) and inter-class (hateful-benign) similarities.

4.6.6 On-the-Fly OoD ICL Evaluation

Hypothesis I (§4.5.4.1) states that the OoD ICL works *when the zero-shot performance is poor*. While this hypothesis provides insights into the role of ICL over the formatting bias mitigation, we acknowledge the significance of selecting the prompting strategy *on-the-fly*; or selecting the strategy *before*

²<https://eval.ai/web/challenges/challenge-page/830/my-submission>

³Random effects vary across samples; fixed effects are constant across all samples.

⁴For example, a fixed effect of LLaVA-1.5 is presupposed to be identical over all the outputs generated by this model.

fully evaluating the zero-shot performance for real-world applications. To enable this real-time selection between zero-shot and OoD ICL prompts without requiring complete in-distribution calibration, we implement a dynamic strategy-switching algorithm. Given the zero-shot test-input representation H_{ZSL} , we train two lightweight auxiliary predictors f_{ZSL} and f_{ICL} to estimate the probability of a correct answer under each strategy. At inference, we choose

$$Y_{\text{alg}} = \begin{cases} L(X_{\text{ICL}}) & \text{if } f_{\text{ICL}}(H_{ZSL}) > f_{ZSL}(H_{ZSL}) \\ L(X_{ZSL}) & \text{otherwise} \end{cases} \quad (4.6)$$

where L is the LLaVA model and X_{ZSL}, X_{ICL} denote the zero-shot and OoD ICL prompts, respectively. Because the decision relies solely on H_{ZSL} , it can be made on-the-fly without full prior ID evaluation.

We instantiate each f as a single linear layer for LLaVA-Llama2 and as a single-layer Transformer for LLaVA-1.5⁵. Auxiliary models are trained on 70% of the GQA and TextVQA test splits (holding out 30% for evaluation), for which we can obtain the binary labels locally. We optimized the auxiliary models using binary cross-entropy loss⁶ with an Adam optimizer [118] for optimization. For LLaVA-Llama2, we upweight the positive (correct) class by a factor of 10 to mitigate label imbalance⁷.

4.7 Results & Discussion

4.7.1 Experiment I: Multi-image Multi-turn OoD

4.7.1.1 Performance Evaluation

Experiment I evaluates the hypothesis that semantically rich M^2 OoD ICL examples induce beneficial representational shifts in LLaVA, leading to improved VQA performance under distributional shifts. First, we compare ID versus one-shot OoD ICL performance to quantify gains attributable to semantically aligned demonstrations (Figure 4.4). On six VQA benchmarks, OoD ICL with LLaVA-Llama2 yields substantial accuracy improvements on

⁵Prior to this on-the-fly evaluation experiment, we performed the qualitative analysis (§4.7.1.2), observing the relatively simple reasoning (boosted task-following ability in ICL) behind LLaVA-Llama2 and more complex findings for LLaVA-1.5. Thus, we assumed the simple mapping between zero-shot and in-context representations for the former, and more complex, potentially non-linear projection for the latter.

⁶<https://docs.pytorch.org/docs/stable/generated/torch.nn.BCELoss.html>

⁷Aside from 1 : 10, We also tested 1 : 1, 1 : 2, 1 : 5, 1 : 20, and 1 : 10 works the best. We also did some preliminary experiments on Focal Loss [119], but did not see a significant performance gain in the preliminary experiments.

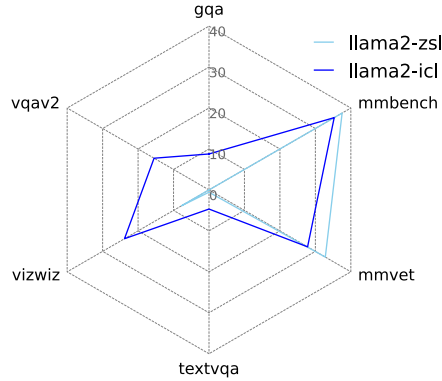


Figure 4.4: Performance summary of LLaVA-Llama2. zsl and icl represent Zero-Shot Learning (ZSL) and Out-of-Distribution In-Context Learning (OoD ICL; denoted simply as *ICL* in the figure legends for brevity). OoD ICL results in better performance for four datasets, where the zero-shot performances are poor.

datasets with relatively low ZSL performances (GQA, VQAv2, VizWiz, and TextVQA) while having minimal impact on relatively-high-ZSL-performance datasets (MMBench and MM-Vet).

4.7.1.2 Qualitative Analysis

Next, to illustrate how OoD ICL examples correct errors and inject missing semantics in practice, we examine representative cases from GQA and TextVQA that we can obtain ground-truth labels locally for precise analysis. Examples demonstrate that semantically relevant OoD prompts selected via CLIP similarity can constrain answer format (e.g. enforcing single-word responses; Figure 4.5) and supply analogical cues (e.g. monetary units; Figure 4.6), while random sampling fails to produce these effects (Figure 4.7). More analysis is in Appendix B.4.2.

4.7.1.3 Representation Analysis

We aim to verify that observed performance gains correlate with measurable shifts in the model’s hidden representations, as predicted by RST. A linear mixed effect model shows that the input-driven representational shift term explains the majority of variance in the ICL-augmented embeddings ($R^2 = 0.59 \pm 0.02$), outperforming the random-effect-only baseline ($R^2 = 0.43 \pm 0.01$), validating the efficiency of our mixed effect model. Next, we analyzed

the regression coefficients of the mixed effect model (Table 4.2). We see that

Table 4.2: Regression Coefficient*100 of the mixed effect model’s prediction with the dummy variables representing the datasets and the models. The prediction shows a much higher coefficient than the dummy variables, validating our models.

variable	coef*100
(Intercept)	9.2 ± 2.1
mm-vet	-0.75 ± 0.7
mmbench	2.81 ± 0.7
textvqa	2.1 ± 0.6
vizwiz	0.16 ± 0.7
vqav2	-0.12 ± 0.6
model	-0.39 ± 0.4
Input Shift	70.33 ± 5.9

the coefficient of the input shift is much higher than that of the variables introduced for the fixed effect, suggesting that semantic content in OoD ICL examples drives the shift under the variances of datasets and models.

4.7.1.4 On-the-Fly OoD ICL Evaluation

Towards real-world applications, we introduce a dynamic selection mechanism (Eq. 4.6) that chooses between ZSL and OoD ICL prompts based solely on zero-shot representation features, eliminating the need for prior ID performance measurements (*on-the-fly*). We show the results in Table 4.3. For LLaVA-Llama2, the on-the-fly strategy’s accuracy is capped by the static M^2 OoD ICL performance, reflecting the model’s extremely low ZSL baseline. This strategy still achieves a 1.3–6.5 point absolute gain over ZSL. In contrast, for LLaVA-1.5, where ZSL and static OoD ICL accuracies are comparable, the on-the-fly approach outperforms both fixed strategies by leveraging their complementary strengths. These results suggest that strategically designed OoD ICL examples, combined with dynamic on-the-fly strategy selection, can effectively mitigate spurious formatting biases to boost semantic reasoning and yield robust performance gains across diverse VQA datasets.

Table 4.3: Accuracy of the on-the-fly context selection. Bold indicates the best performance in each row. In the LLaVA-Llama2 case, the performance is bounded by the OoD ICL (as with the figures, denoted as *ICL* in the table legends) accuracy since ZSL performance is extremely low. In the LLaVA-1.5 case, it outperforms *both ZSL and ICL*, suggesting its efficiency when the performance of the two strategies is comparable.

Model	Dataset	ZSL	ICL	Eq. 4.6
LLaVA-Llama2	TextVQA	0.9	4.7	2.2
	GQA	0.0	9.0	6.5
LLaVA-1.5	TextVQA	61.6	57.0	63.8
	GQA	65.7	56.0	68.2

4.7.2 Experiment II: Single-image Multi-turn OoD

4.7.2.1 Note on notation in Experiment II

In Experiment II, although SM OoD can be seen as a variant of ICL that the model generates an ICL example for self-improvement, we use ICL to refer to the M^2 OoD approach, and CFP to refer to the SM OoD approach. This is because M^2 OoD is a typical form of ICL where the model is provided with an example given by the user (in this case an image and a caption) to compare against the test input. To avoid confusion, we repeat this notation for multiple times in this section.

4.7.2.2 Performance Evaluation

Experiment II tests whether the SM OoD (CFP) approach mitigates textual biases in a text-dominant VQA setting. Specifically, we apply CFP to generate benign examples and compare model predictions on the HMC with ZSL and M^2 OoD conditions (Table 4.4). In Experiment II, we use M^2 OoD and ICL interchangeably. Supporting Hypothesis I, CFP approach improves F1 from 61.4% (ZSL) to 62.2%, whereas ICL decreases performance to 58.5%.

Table 4.4: Hateful memes detection performance. CFP’s performance is better than ZSL while regular ICL (M^2 OoD) dropped the performance, supporting Hypothesis II.

setting	ZSL	ICL	CFP
f1*100	61.4 ± 0.5	58.5 ± 0.9	62.2 ± 0.3

4.7.2.3 Representation Analysis

We investigate how CFP prompts reshape the representation space to separate hateful and benign classes, reducing label confounding (Figure 4.8). Representational similarity matrices show that CFP substantially lowers cross-label similarity, effectively disentangling class representations and attenuating textual shortcuts. These results suggest that CFP not only boosts F1 beyond both ZSL and M^2 OoD ICL baselines but also markedly enhances the latent separation between hateful and benign representations in the presence of both textual and formatting biases.

4.7.3 Discussion

The core premise of RST is that ICL elicits two distinct transformations in a model’s hidden states: an *input shift* driven by semantic content, and an *output shift* reflecting downstream adaptation. Under RST, an effective OoD generalization requires amplifying the dynamic semantic component of the input shift relative to the static *formatting bias* term.

Experiment I validates this principle: when zero-shot semantics are weak, M^2 OoD ICL demonstrations induce larger semantic shifts, and these shifts are translated into accuracy gains over ZSL (Figure 4.4). Qualitative case studies further confirm that CLIP-guided example retrieval yields more pronounced semantic shifts than random sampling (§4.7.1.2). The on-the-fly strategy leverages early zero-shot representations to choose between ZSL and OoD prompts, matching static OoD gains for weaker models and exceeding both strategies for stronger models by dynamically harnessing their complementary advantages (Table 4.3).

Experiment II extends RST to single-image settings: CFP (SM OoD) not only improves performance (Table 4.4) but also produces a marked drop in inter-class cosine similarity of the input representation (Figure 4.8), indicating clearer separation of hateful versus benign representations. This demonstrates that even in text-dominant tasks, carefully designed OoD examples can amplify the semantic shift and mitigate label-specific textual

biases and formatting biases.

Although this work focuses on proposing a theory about the role of ICL on formatting bias mitigation, our experiments showed the practical guidelines for selecting ICL examples, particularly in challenging OoD problems. Under the necessities of multi-image or multi-turn reasoning in the absence of models that are not trained on such queries, the practitioners can either (1) add the meaningful context, such as CLIP-based ICL examples despite the formatting variance from the training data, or (2) utilize the model’s ability to curate ICL examples.

Together, these findings offer evidence-grounded support for RST’s view that the representational impact of ICL can and should be dissected into semantic and format contributions.

Previous efforts on building interpretability theories for ICL have validated the concept of meta-gradient, attention weight used as a form of gradient [96, 97]. Meta-gradient backbones RST, which provides an analytical framework for formatting bias mitigation. Towards separating the semantic information from formatting bias, interpretability studies disentangled a few aspects of the semantics [84–86, 88]. Beyond the limited aspects of the semantics and semantics-format *separation*, RST provides a unifying framework for *mitigating* the formatting bias for amplifying the semantic understanding of the VLMs.

Although our theory provided the general framework for ICL analysis, its detailed mechanism is elusive. For example, linguistic patterns inherent in ICL examples may contribute to OoD ICL; as an explanation using parallel structures [104], or the repetitive occurrence of specific tokens, the model may learn the repetitive occurrences of the image token inserted before textual context ($[Image\ token] \rightarrow [Textual\ context] \rightarrow [Image\ token] \rightarrow [Test\ input]$). Likewise, the model’s capability of learning such a token sequence is a promising candidate for the mechanism underlying the representational shift. One of the famous mechanisms is an induction head [101, 102], with which the model performs token completion. In ICL for labeled datasets, the induction head may utilize the label to perform a simple completion like $[ICL\ example\ A] \rightarrow [Label\ Y_1] \rightarrow [Test\ input\ B, which\ is\ similar\ to\ A] \rightarrow [Completed\ Label\ Y_1]$. While this simple completion is less likely in unlabeled datasets (e.g. MM-Vet), pattern analysis is a promising area of research. Also, our results support the presence of function vectors [103] in that the latent space right after presenting an ICL example influences model behavior. As with preceding studies, we plan to explore which attention heads cause the representational shift in future work.

Another fundamental topic is the definition of ICL itself. The original paper [12] defines ICL as “a learning method using the input of a pretrained

*language model as a form of task specification: the model is conditioned on a natural language instruction and/or a few demonstrations of the task and is then expected to complete further instances of the task simply by predicting what comes next*⁸. Although our qualitative analysis suggests that task comprehension is improved by the proposed method (Figure 4.5), the mechanism behind the performance gain may be slightly different from the explanation above, since LLaVA variants are *not* conditioned by multiple pairs of images in OoD ICL. Since the mechanism of ICL in general is still an open question, future work may provide a unified explanation of, or distinction between, the traditional ICL and the phenomenon studied in this thesis.

4.7.4 Limitations

While our study provides valuable insights into formatting bias mitigation, there are several limitations and future research directions that warrant further investigation.

We limited our evaluation to one-shot ICL because few-shot ICL introduces additional complexities to our analysis, such as the fixed or random effects of varying the number of images. However, we acknowledge that applicability to few-shot ICL is critical, and we plan to tackle this challenge in future work.

Although RST can be used to analyze arbitrary problems, the largest limitation for the time being is its generalizability; to foresee the performance improvement in another problem, we need another hypothesis tailored to that problem. Towards the automatic formulation of the novel hypothesis, we believe the flexibility of semantic and formatting bias terms (Eq. B.4) is the key. This study is also limited linguistically; we only used English datasets.

Another empirical limitation is oversimplicity; we acknowledge that fully decomposing input information into *distinct* format and semantic components is challenging. In other words, formatting bias and statistics may *interact* in the real world. For example, certain semantic content may only be understandable within specific formats. A complicated mathematical concept, for instance, might be best comprehended through equations. We acknowledge this interplay and will consider it in future work.

From a theoretical point of view, we have an intuitive leap from the existing works on meta-gradient; a nonlinearity [96, 97]. Despite previous works on *secretly* linear nature of a nonlinear Transformer [120] and our empirical

⁸Slightly modified to fit in the context of this thesis

findings supporting RST, applying the concept developed on a linear variant to the nonlinear one might hinder the precise evaluation. Recently, [94] proposed a theory for the nonlinear Transformer variants with the help of contrastive learning [121]. Unifying RST with their approach might provide a robust theoretical grounding.

In addition, whether the input shift *causes* the output shift is still elusive. An approach is to hire an MI method, such as path patching [122, 123]. Training phase mechanisms such as grokking or double descent [124] should also provide an explanation for the mechanism behind representational shift.

4.8 Conclusion

In this chapter, we tackled formatting bias, where non-semantic constraints such as limited image counts or conversation turns skew VLM reasoning, by operationalizing a theoretical approach and crafting targeted OoD ICL interventions. Through M^2 OoD ICL demonstrations, Experiment I overcame single-image/single-turn restrictions to yield accuracy gains on low-zero-shot-performance VQA tasks. Experiment II employed single-image CFP (SM OoD ICL) to neutralize turn-count limitations, boosting HMD performance and markedly improving latent class separability. Together, these findings show that amplifying dynamic semantic shifts relative to static formatting factors can effectively neutralize format-induced artifacts. Future work will investigate automatic problem-specific hypothesis curation based on RST, RepE for driving the representational shifts towards desired directions, and extensions of RST to pure LLM settings.

*Incorrect ZSL response by
LLaVA-Llama2.*

User:



Are there drapes to the right of the bed?

Answer the question using a single word or phrase.

Assistant:

yes, there are drapes to the right of the bed

Correct ICL response by LLaVA-Llama2

User:



Do you see sheets on the bed?

Answer the question using a single word or phrase.

Assistant:

no

User:



Are there drapes to the right of the bed?

Answer the question using a single word or phrase.

Assistant:

yes

Figure 4.5: A representative LLaVA-Llama2’s response from the GQA dataset. (A) LLaVA-Llama2 does not follow a part of the instruction where it is required to answer the question in a single word. (B) LLaVA-Llama2 responded in a single word with an ICL example.

Incorrect ZSL response by LLaVA-1.5.

User:



How much is the copper colored coin worth?

Reference OCR token: [...], ONE, 2aH

Answer the question using a single word or phrase.

Assistant:

1

Correct ICL response by LLaVA-1.5.

User:



Provide a one-sentence caption for the provided image.

Reference OCR token: L

Assistant:

A copper coin featuring the words ONE PENNY sits on top of a white surface.

User:



How much is the copper colored coin worth?

Reference OCR token: [...], ONE, 2aH

Answer the question using a single word or phrase.

Assistant:

1 penny

Figure 4.6: A representative LLaVA-1.5’s response for which OoD ICL impacts the performance **positively** from the TextVQA dataset. (A) LLaVA-1.5 answered the monetary value question without its unit (*just 1, without penny*), potentially due to the missing OCR token for the unit. (B) LLaVA-1.5 answered **correctly**, potentially because the ICL example attributed **one penny** to the value of the copper coin.

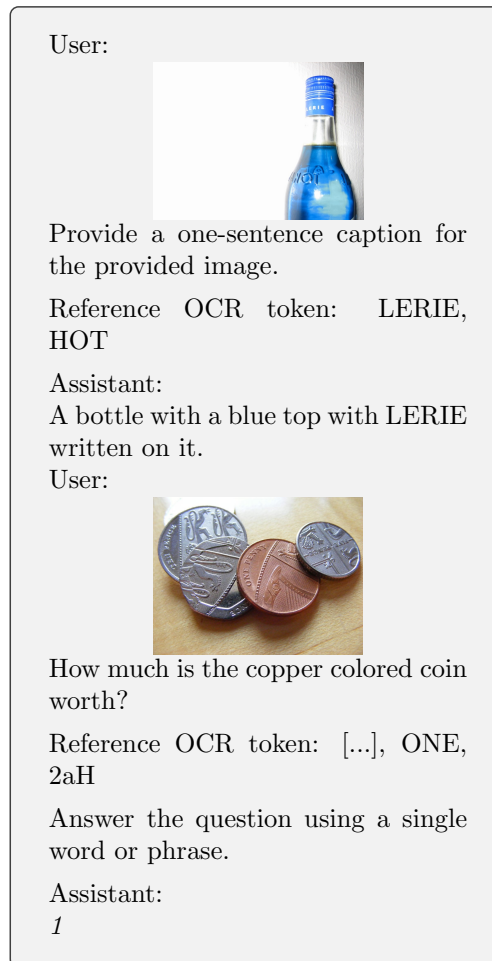


Figure 4.7: A representative LLaVA-1.5’s response when the ICL example is randomly sampled from the training dataset. Seemingly irrelevant image-text ICL example does not affect the model’s response.

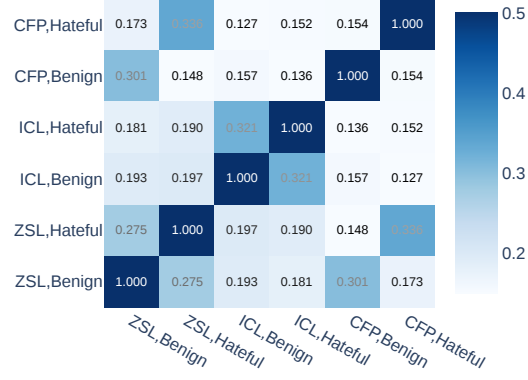


Figure 4.8: Representational shift across the learning type. Each entry is the similarity of the input between two conditions. For example, the left-top value 0.173 is the similarity of the input between hateful samples of a CFP setting and benign samples of a ZSL setting. While the hateful samples and the benign samples are similar for ZSL and ICL (i.e., M^2 OoD) settings, CFP hateful samples and benign samples are less similar.

Chapter 5

Conclusion

5.1 Are the two studies disjoint?

In this dissertation, we studied two types of presentation biases: modality bias (Chapter 3) and formatting bias (Chapter 4). Although the two studies share a high-level research interest in interpreting/mitigating individual presentation bias, the broad scope may have resulted in discrepancies of the methods (causality metrics for modality bias/theory-driven context selection for formatting bias) and of the problems (HM/VQA). However, I argue that the two chapters are closely connected within these layers.

First, the two methods are compatible; while the *bottom-up*¹ approach, such as attribution metrics, is grounded on empirical evidence, the downside of this approach is its generalizability to other datasets/models/settings than those studied in a specific study. In contrast, although the *top-down* approach, such as theories, provides a setting-agnostic general framework, it often struggles with applying to real-world complexities. A discussion of combining the two type of methods is in 5.4. For more discussion on bottom-up/top-down approaches, I refer the readers to the recent review on MI [125].

Second, HM is a special case of VQA in the era of generative VLMs (that is, VLMs can perform the HMD task when they are asked to answer whether the given image-caption pair is hateful or not; see §4.7.2). In Chapter 3, I focused on HMC to assess the impact of the causal graph representing its data generation process (§3.5.3). Although this approach allows me to logically ground the analysis, this also limits the generalizability of the proposed method (§3.7.5). To address this limitation, I used general VQA datasets in Chapter 4.

¹Finding the insights from each of the parameters

5.2 Contribution Summary

In this dissertation, I analyzed the presentation biases *within* the VLMs for HMD and VQA, and proposed debiasing methods based on the insights obtained from their inner workings.

First in Chapter 3, we proposed metrics for quantifying the modality bias in predicted hatefulness (MITE) and in attention matrices (MIDAS). The analysis of the three BERT-based HMD models shows the predominant biases towards textual information, especially in VisualBERT. The calibration of the prediction by MIDAS suggests that this metric could be utilized for debiasing the HMD models, especially when the baseline classifier is biased. Next in Chapter 4, we proposed RST, a theory about the role of ICL on formatting bias mitigation, or how a VLM comprehends semantically meaningful information in the presence of the formatting bias. Based on this theory, we proposed two hypotheses for *how* and *when* ICL works in the challenging OoD settings, where the model needs to understand the semantics in the presence of heavy formatting bias (the unseen number of images and/or the unseen number of conversation turns). The empirical results on six VQA datasets and HMC dataset validate RST and the two hypotheses.

In summary, a high-level contribution of this dissertation is summarized as:

- **High-Level Analyses of VLM’s Inner Workings for Presentation Bias Detection and Mitigation.** Beyond the existing works on the outcomes (predicted probabilities and text-formed responses), this dissertation provided metrics for the modality bias within a VLM (Chapter 3) and theoretical accounts of how a VLM addresses formatting bias (Chapter 4).
- **Mitigating Presentation Biases upon Model Internals.** Whereas the previous works on interpretability focused on understanding the model internals, this dissertation provided the practical guidelines for improving the base VLMs’ decisions based on the understanding.

5.3 Limitations

In this dissertation, I proposed diverse solutions to different problems (two metrics for modality bias and an ICL theory for formatting bias) under the same philosophy. Therefore, a major limitation of this study is the scalability of the findings; Whether the proposed frameworks can be used for different problems is unclear. To this end, unifying both approaches (e.g. an automatically generated bias metric based on the theoretical grounding)

is a promising area of study.

Another limitation is in its generalizability, especially for HMD; applicability to other datasets (e.g. MUTE [77] for HMD) is important. Additionally, this dissertation explored a limited range of VLMs (three BERT-based HMD models and LLaVA). A comprehensive analysis of the different models, such as the emerging reasoning models like DeepSeek [126], is critical for better insights.

As mentioned in §4.7.4, the gap between the linearity-presupposing theoretical literature and the non-linearity inherent in the real-world VLM applications is challenging. Additionally to the meta-gradient framework for a regular Transformers [94], expanding the scope to a broader range of theories, such as a singular learning theory [127, 128] and an infra-Bayesianism [129] is a promising avenue.

5.4 Future Work

As with Chapter 1, the ultimate goal of interpretability is the human-VLM alignment. To fill in the gap between interpretability and alignment, future work should involve more interventional efforts; as we isolated critical mechanisms within VLMs, the next step is to monitor their test-time behaviors and to *correct* the unexpected ones. MIDAS-based HMD calibration (§3.6.3 and 3.7.3 in Chapter 3) and on-the-fly ICL evaluation (§4.6.6 and 4.7.1.4 in Chapter 4) show promising results; yet the application and performance gains are limited. Taking a hateful meme detector (Chapter 3) as an example, a high text-to-text MIDAS score for an image-oriented task suggests the presence of the pertaining bias, or shortcut learning [130]. In that case, the fine-grained interventions in the model (in the attention matrix, linear projections, etc.) may alleviate the unwanted behavior while keeping the inherently unbiased decisions intact. For more discussion, I refer the readers to the comprehensive review of MI [125] and RepE [63].

Another problem that interventional techniques may address is the causal role of the bias detected in visualization over the outcome (predicted probabilities and text-form responses). For example, the overestimated attribution of attention to textual tokens can be suppressed by attribution patching [131] to test its causal impact on prediction and performance. Towards the real-world applications, an important topic is the discussion about potential use cases. In general, while the gradient-based XAI approaches (Chapter 3) provide supporting evidence for domain experts and decision makers, theoretical understanding of the mechanism underlying the complex model (Chapter 4) benefits researchers and engineers more. Therefore, the real-world applica-

tion would be multi-layered; for example, while showing an attribution metric to the decision makers, the engineers can investigate the potential cause of its value.

Additionally, I should perform a more comprehensive analysis on the cross-factor interactions. For example, our causality framework in Chapter 3 did not take into account the presence of unimodal hatefulness (§3.7.5) and other factors such as the annotator’s preference².

5.5 Concluding Remarks

In this dissertation, I analyzed the presentation biases within VLMs for HMD and VQA, and proposed debiasing methods based on the insights obtained from their inner workings. The proposed metrics for modality bias (MITE and MIDAS) and the theory of formatting bias (RST) provide a new perspective on the interpretability of VLMs, which can be utilized for improving their performance in real-world applications. I hope that the findings of this dissertation contribute to the ongoing efforts in making VLMs more interpretable and aligned with human values.

²Note that curation processes of the most hate speech or hateful meme datasets, including HMC, involve cross-annotator agreements, which is expected to minimize the impact of a single annotator’s decisions. For more details, I refer the readers to the original HMC publication [18]

Appendix A

Appendix for Hateful Meme Detection

A.1 Formal Relation between miATE and MIDAS

In this section, we investigate the formal relationship between MITE and MIDAS. Although MITE and MIDAS operate on different levels, MITE on the model’s output probabilities and MIDAS on attention-weight attributions, they are closely related. Concretely, if we write the one-step gradient of the output θ with respect to the attention tensor A as

$$G(A) = \frac{\partial \theta(A)}{\partial A} \quad (\text{A.1})$$

then, for a single input (T, I) , the integrated-gradient attribution satisfies (when $\alpha = 1$):

$$\text{attr}^{(T,I)} \approx A \odot G(A) \quad (\text{A.2})$$

Now, by definition,

$$\text{miATE} = \theta(A^{(T_1, I_1)}) - \left[\theta(A^{(T_1, I_0)}) + \theta(A^{(T_0, I_1)}) \right] \quad (\text{A.3})$$

If we perform a first-order Taylor expansion of $\theta(\cdot)$ around a *baseline* attention matrix $A^{(T_0, I_0)}$, then

$$\theta(A^{(T_1, I_1)}) \approx \theta(A^{(T_0, I_0)}) + \langle G(A^{(T_0, I_0)}), A^{(T_1, I_1)} - A^{(T_0, I_0)} \rangle \quad (\text{A.4})$$

and similarly for $\theta(A^{(T_1, I_0)})$ and $\theta(A^{(T_0, I_1)})$. Subtracting yields

$$\text{miATE} \approx \left\langle G(A^{(T_0, I_0)}), A^{(T_1, I_1)} - (A^{(T_1, I_0)} + A^{(T_0, I_1)}) \right\rangle \quad (\text{A.5})$$

Observe that

$$A^{(T_1, I_1)} - (A^{(T_1, I_0)} + A^{(T_0, I_1)}) \quad (\text{A.6})$$

is roughly the *change* in attention-weights that occurs only when both modalities jointly signal hate (the same *difference term* that appears inside MIDAS). Since MIDAS_{it} is essentially the *average* of

$$A_{i,j}^{(T_1, I_1)} = (A_{i,j}^{(T_1, I_0)} + A_{i,j}^{(T_0, I_1)}) \quad (\text{A.7})$$

over all $(i, j) \in \mathcal{P}_t$, we see that

$$\text{miATE} \approx \sum_{t \in \{\text{text}, \text{image}, \text{cross}\}} \text{MIDAS}_t \times \sum_{(i,j) \in \mathcal{P}_t} N(G_{i,j}) \quad (\text{A.8})$$

where $N(\cdot)$ is a monotonic normalization of gradients to $[0, 1]$. Hence, *miATE* (the output-level causal effect) can be viewed as a *gradient-weighted sum* of the three MIDAS components. In practice, we do not compute this summation explicitly; rather, we use MITE and MIDAS as two complementary diagnostics:

- *miATE* tells us whether θ 's *predictions* depend on cross-modal synergy versus unimodal shortcuts.
- MIDAS tells us *which type* of attention patterns (text, image, or cross) the model is using (and how much each pattern *jumps* when both modalities truly signal hate).

Because MIDAS_t is defined per modality-interaction category and is inexpensive to compute (once we have $\text{attr}(\cdot)$), it serves as a fine-grained, layer-wise proxy for the global *miATE*. Empirically, we find that as $\text{MIDAS}_{\text{text}}$ grows (relative to $\text{MIDAS}_{\text{cross}}$), the corresponding $\text{miATE}_{\text{T2T}} - \text{miATE}_{\text{I2I}}$ gap also widens (i.e., the model leans more heavily on text).

A.2 Further Exploration for Local Explainability

Sample analysis for Oscar (fig: A.1) shows a similar trend to UNITER (fig: 3.4). Interestingly, VisualBERT (fig: A.2) does not attend to the key components (woman or cargo) in the image, supporting its bias towards textual information.

A.3 Breakdown of Attention Attribution Score

attr (3.4) is the product of the attention weight matrix and the integral of the gradient. To see the separate impact, we replaced the attr term of the

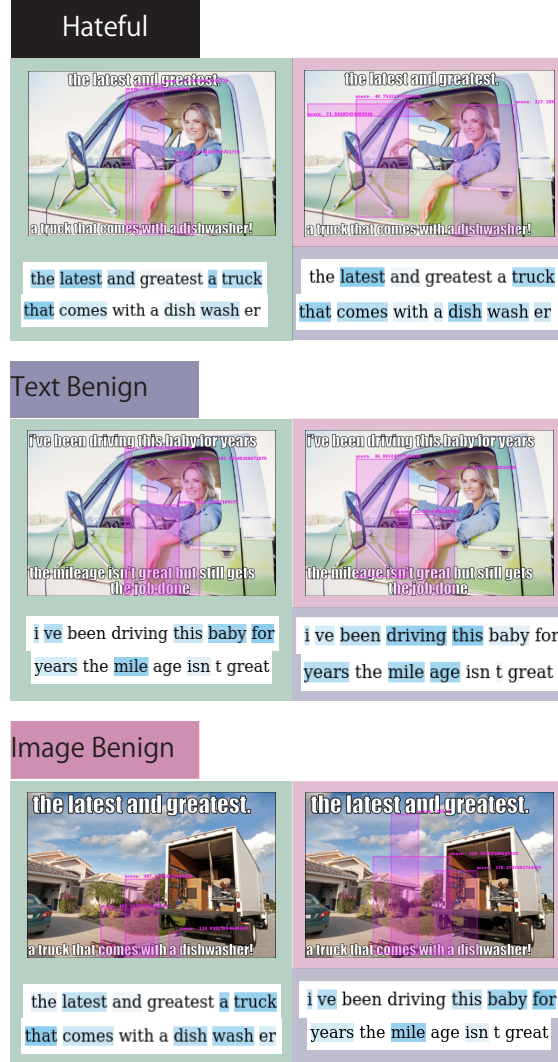


Figure A.1: Sampled derived from Oscar.

$MIDAS$ equation (Eq. 3.7) with the attention $MIDAS_{att}$ or the gradient $MIDAS_{grad}$ for comparison. In general, $MIDAS_{att}$ (fig: A.3-5) shows a more similar trend to the original $MIDAS$ than $MIDAS_{grad}$ (fig: A.6-8). This result implies that the attention weights decide the model's strategy, while the gradient adjusts the impact of the individual component.



Figure A.2: Samples derived from VisualBERT.

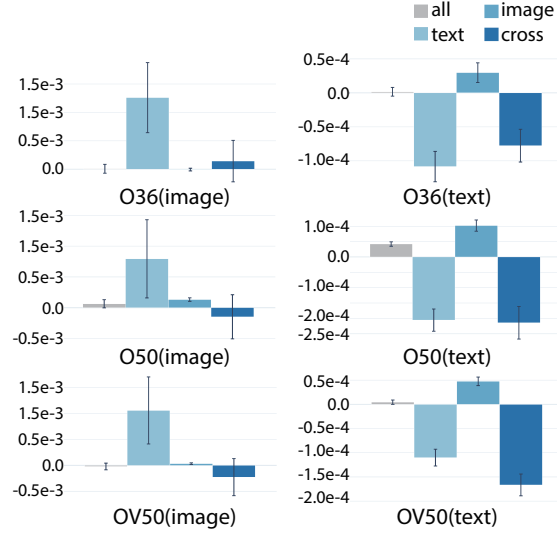


Figure A.3: Oscar $MIDAS_{att}$.

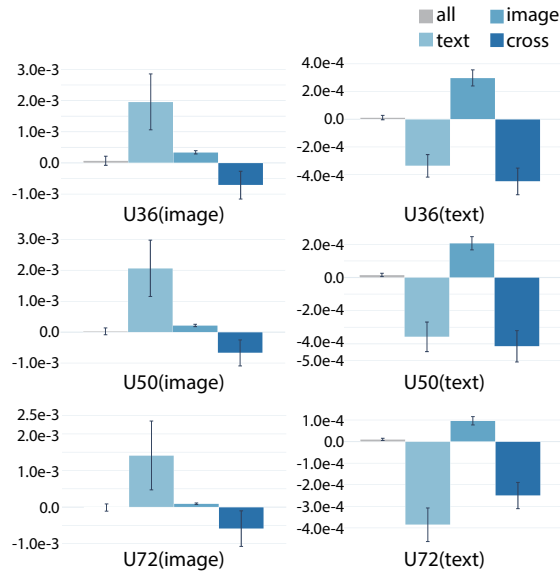


Figure A.4: UNITER $MIDAS_{att}$.

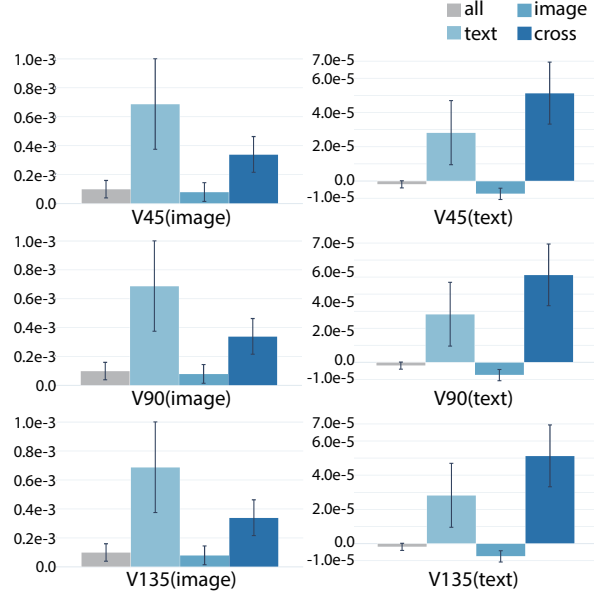


Figure A.5: VisusalBERT $MIDAS_{att}$.

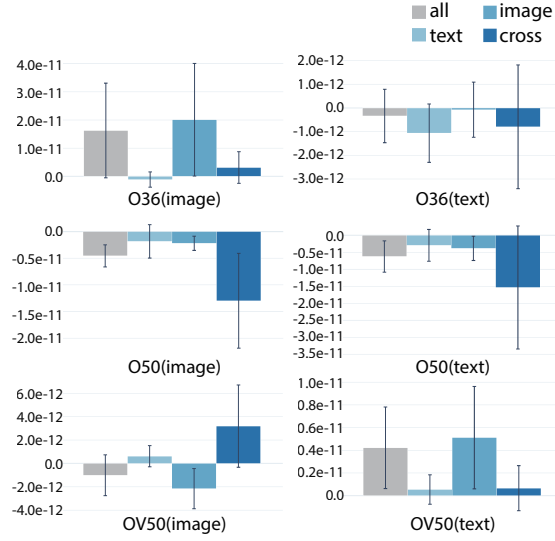


Figure A.6: Oscar $MIDAS_{grad}$.

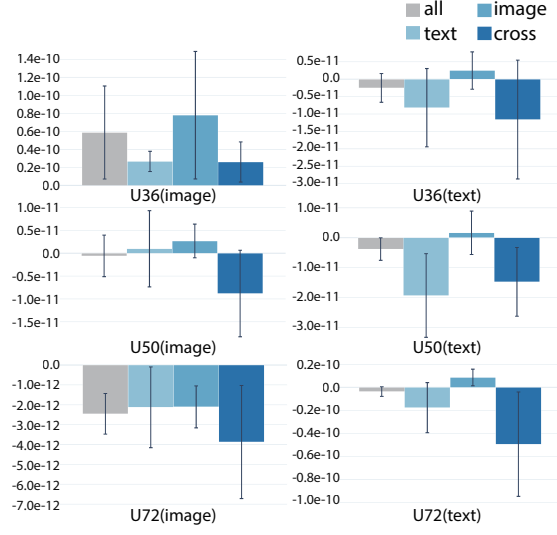


Figure A.7: UNITER $MIDAS_{grad}$.

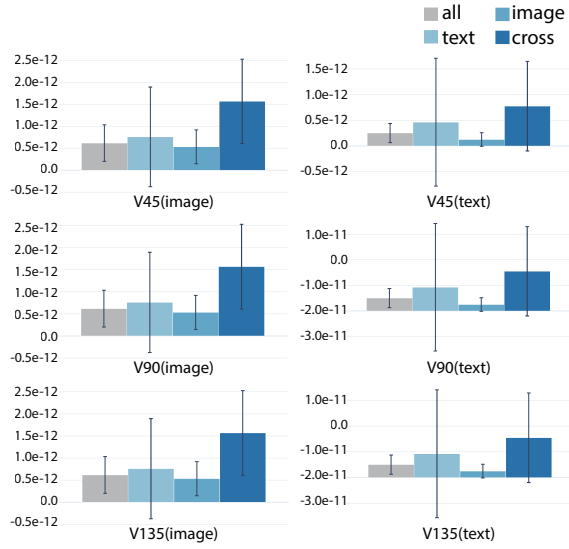


Figure A.8: VisualBERT $MIDAS_{grad}$.

Appendix B

Appendix for Out-of-Distribution Generalization

B.1 Formalization of Representational Shift Theory

B.1.1 Representational Shift

We formalize Representational Shift Theory (RST) by analyzing the difference between the zero-shot input-output pair $\{H_{zsl}, Y_{zsl}\}$ and that of ICL $\{H_{icl}, Y_{icl}\}$. Assuming that the effect of the instruction over an ICL example and over a zero-shot input is identical, i.e., $\Delta W_{inst/zsl} \simeq \Delta W_{inst/icl}$, we obtain the input shift:

$$H_{icl} - H_{zsl} \simeq -\Delta W_{icl/zsl} H_{inst} \quad (\text{B.1})$$

Applying this to the output, we observe an output shift:

$$Y_{icl} - Y_{zsl} = -W_{emb} \Delta W_{icl/zsl} H_{inst} \quad (\text{B.2})$$

Equations B.1 and B.2 represent the basic concept of RST. Note that the LLM’s final output is a sequence of words, but we use the representation of the last decoder layer as the output for analysis. To intuitively analyze the multi-dimensional representation, we use a distance metric $D_{X/Y} \propto X - Y$:

$$D_{Y_{icl}/Y_{zsl}} = W_{RST} D_{H_{icl}/H_{zsl}} \quad (\text{B.3})$$

where $W_{RST} = -H_{inst}^T W_{emb}$

In practice, we use cosine similarity as the distance metric. This formalization allows us to analyze the effect of ICL by comparing the distances between representations and outputs.

B.1.2 Formatting Bias Mitigation

To disentangle semantics from formatting bias, we assume that the two concepts are independent. In RST, this implies that the weight updates

due to semantics ΔW^{sem} and due to formats ΔW^{stat} are discernible. We suggest that the semantic distance D^{sem} and the formatting distance D^{stat} are also separable, as indicated by the relevance of representational shift and the distance metric (Equation B.3). We formalize the disentanglement as:

$$\begin{aligned}\Delta W_{icl/zsl} &= \Delta W_{icl/zsl}^{sem} + \Delta W_{icl/zsl}^{stat} \\ D_{H_{icl}/H_{zsl}} &= D_{H_{icl}/H_{zsl}}^{sem} + D_{H_{icl}/H_{zsl}}^{stat}\end{aligned}\tag{B.4}$$

This separation allows us to analyze how semantics and formatting biases individually contribute to the representational shift.

B.1.3 OoD Generalization as Formatting Bias Mitigation

An OoD input forces an LLM to generalize the same semantics under a significant distributional difference in formats. Since the format difference is consistent across all test inputs, its effect on the representational shift is constant (*fixed* effect). In contrast, the semantic term’s effect varies across samples (*random* effect). Under this assumption, we formalize OoD generalization as a mixed effect:

$$D_{Y_{icl}/Y_{zsl}} = W_{RST}(D_{W_{icl}/W_{zsl}}^{sem} + W^{stat})\tag{B.5}$$

B.1.3.1 Hypothesis I: MM OoD

Our first hypothesis is that MM OoD ICL examples are effective when the zero-shot input does not provide enough semantics to the model (i.e., poor zero-shot performance):

$$\begin{aligned}D_{W_{icl}/W_{zsl}}^{sem} &= W_{icl}^{sem} - W_{zsl}^{sem} \\ D_{Y_{icl}/Y_{zsl}} &= W_{RST}(W_{icl}^{sem} + W^{stat}) \\ \text{where } W_{zsl}^{sem} &\ll W_{icl}^{sem}\end{aligned}\tag{B.6}$$

One scenario is the lack of regularization in the attention matrix. If semantically similar ICL examples amplify the relevant context, our approach can alleviate irrelevant context, improving performance.

B.1.3.2 Hypothesis II: SM OoD

When textual semantics $W^{sem}(T)$ are more informative than image semantics $W^{sem}(I)$, enhancing the textual term through SM OoD ICL examples can

be beneficial:

$$\begin{aligned} W_{icl}^{sem} &= W_{icl}^{sem}(T) + W_{icl}^{sem}(I) \\ D_{Y_{icl}/Y_{zsl}} &= W_{RST}(W_{icl}^{sem}(T) + W^{stat}) \\ \text{where } \Delta W_{icl/zsl}^{sem}(T) &+ \Delta W_{icl/zsl}^{sem}(I) \end{aligned} \quad (\text{B.7})$$

For brevity, we assume the independence of semantics over the two modalities. This scenario is effective in addressing label bias [132], where the model’s prediction may be biased toward certain labels due to over-reliance on input formats.

B.2 Other Formalization

B.2.1 Mixed Effect Model

In Experiment I, we implemented a linear mixed effect model to analyze the mixed effect of the input shift and confounding variables over the output shift. The model predicts the shifted representation \hat{H}_{icl} as:

$$\hat{H}_{icl} = (W_r + W_f I) H_{zsl} + W_0 \quad (\text{B.8})$$

Here, W_r represents the random effect, W_f represents the fixed effect, I is the embedding of fixed components (dataset and model), and W_0 is a bias term. The baseline model includes only the random effect:

$$\hat{H}_{icl} = W_{random} H_{zsl} + W_0 \quad (\text{B.9})$$

By comparing the performance of these models, we assessed the contributions of the random and fixed effects.

B.2.2 Representational Analysis

In Experiment II, to visualize label bias, we estimated the input shift weight W_{RST} for different conditions (ID, M^2 OoD, SM OoD) and labels (benign, hateful). We computed the cosine similarity between weights to analyze how different approaches affect the model’s internal representations:

$$\begin{bmatrix} \text{sim}(W_0^{zsl}, W_0^{zsl}) & \cdots & \text{sim}(W_0^{zsl}, W_1^{cfp}) \\ \vdots & \ddots & \vdots \\ \text{sim}(W_1^{cfp}, W_0^{zsl}) & \cdots & \text{sim}(W_1^{cfp}, W_1^{cfp}) \end{bmatrix} \quad (\text{B.10})$$

Lower cross-label similarity indicates that the model better distinguishes between classes, reducing label bias. Note that we do not explicitly consider

layer normalization in our formulation (§4.5.2.2 and B.1.1), our cosine similarity-based analysis implicitly accounts for normalization, focusing on direction rather than absolute values.

B.3 Implementation Details

Experiments are conducted on a single NVIDIA A100 80GB GPU with Linux OS. Unless stated otherwise, all codes are written in Python 3.9. Statistical arguments are based on a t-test and bootstrapping with 1,000 resamples. We run the models once with a random seed of 1987. Eq. B.8 and Eq. B.9 are implemented on a PyTorch backend¹ and trained to maximize the cosine similarity of the output shift with Pytorch Metric Learning package²’s SelfSupervisedLoss under the AdamW optimizer [133]. We extract 1,000 samples from each dataset and hold out 20% as a test set. The performance of this mixed effect model is evaluated using the marginal/conditional R^2 [134]. To maintain the experiment’s integrity while utilizing a wide range of statistical tools, the R language’s *lmer* package is called from the Python environment via *rpy2*³ module.

B.4 Additional Results

B.4.1 Performance Summary of LLaVA-1.5

We show the performance summary of LLaVA-1.5 (Figure B.1). This variant outperforms LLaVA-Llama2 (Figure 4.4) for all the VQA datasets we used, reflecting the authors’ additional efforts. In contrast with LLaVA-Llama2, OoD ICL examples do not contribute to performance gain in LLaVA-1.5. These results further support Hypothesis I (§4.5.4.1), which states that the M^2 OoD works *when the ID performance is poor*.

B.4.2 Additional Analysis on Qualitative Results

In Figure 4.5, we suggested the enhanced task-following capabilities of LLaVA-Llama2 via OoD ICL. To investigate this finding further, we measured the number of words of the correct and incorrect OoD ICL responses

¹<https://pytorch.org/>

²<https://kevinmusgrave.github.io/pytorch-metric-learning/>

³<https://rpy2.github.io/doc.html>

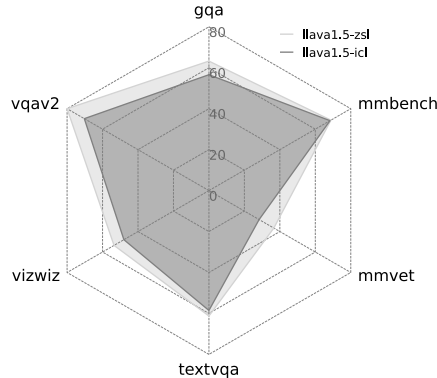


Figure B.1: The performance summary of LLaVA-1.5. OoD ICL dropped the performance, suggesting the rich semantics in the test input.

Table B.1: Distribution of response lengths for correct vs. incorrect predictions made by LLaVA-Llama2 for the GQA dataset. Note the high occurrence of single-word responses in correct predictions.

Prediction	Response Length	
	More Than Two	Single Words
Correct	40	1,064
Incorrect	10,208	1,262

(Table B.1). We observed a higher single-word response rate in correct responses, supporting the qualitative result.

To get further insights about the role of OoD ICL over the VQA performance, we show a *negative* example from LLaVA-1.5’s responses (Figure B.2). We observe the wrong predicted labels (*pine* and *belt*) in the ICL example in this case study, which the model may refer to in making the *incorrect* prediction.

B.4.3 Mixed Effect of Semantics and Formatting Bias: An Arbitrary Argument?

Although we believe that our assumptions for using a mixed effect model (Appendix B.2.1) in Experiment I is logically sound, we acknowledge that this model choice may be arbitrary. For example, we can also model the *random* effect of a *formatting* pattern—For example, the potentially negative impact

Table B.2: Weight means and standard deviations ($\times 10\,000$) for random-effect only model vs. mixed effect model.

Weight Type	Mean \pm Std ($\times 10^4$)	
	Random-Only	Mixed
Slope	1.88 ± 0.09	110.23 ± 2.75
Bias	-6.33 ± 0.38	-508.79 ± 27.07

of an OCR tag present in the ICL example (Fig. B.2) may be considered as such. One hypothesis is that the random effect of the semantics impacts the prediction *positively*, while the fixed effect of the formatting bias works *negatively*. To test this idea, we analyzed the weights of the mixed effect model (Eq. B.8) and the random-effect-only baseline (Eq. B.9). We show the result in Table B.2. Compared with the baseline (left), we can see that the mixed effect model (right) has a larger slope term and a smaller bias term in average. We take this as evidence supporting our hypothesis.

B.4.4 High-Level Analysis on Mixed Effect

In addition to fine-grained analysis in Table 4.2, we analyzed the dataset-level mixed effect. In this analysis, the effects are represented as a coefficient of the corresponding one-hot encodings. Specifically, we modeled the accuracy of each dataset as a sum of the effect of a variable representing the presence/absence of an OoD ICL example and that of the variable representing the models and datasets. The result suggests that the model variable drives the explanatory power at this level, consistent with the performance summary (Fig. 4.4 and B.1), which shows the drastic improvement of LLaVA-1.5 over LLaVA-Llama2.

B.4.5 Preliminary ID Analysis: InternVL

To test if the findings about LLaVA is transferred to an ID setting, we also use InternVL (1-2 billion) for its limited ⁴ yet tested multi-image capabilities by multi-image datasets like MMMU [135].

In the case of InternVL, MM OoD generally dropped the performance, potentially because of its high performance and multi-image resource shortage (Fig. B.3).

⁴<https://github.com/OpenGVLab/InternVL/issues/419>

Variable		R ² *100	
Fixed	Random	Fixed	Random
model	model	22.6 ± 3.0	52.0 ± 8.8
dataset	ICL	0.3 ± 0.1	0.5 ± 0.2
model	ICL	33.5 ± 2.4	33.6 ± 2.5
dataset	model	0.2 ± 0.1	49.5 ± 2.7
all	all	23.7 ± 4.4	53.7 ± 8.8

Table B.3: Regression coefficients of the variables representing model (LLaVA 1.5 or LLaVA-Llama2), dataset, and presence/absence of ICL examples. *all* represents the result of an all-variable model. R^2 values are multiplied by 100 for brevity. The result only with the model variable is similar to the all-variable model, reflecting the performance gap between LLaVA 1.5 and LLaVA-Llama2 (Fig. 4.4 and B.1).

To see whether the task difficulty (i.e., semantic poorness to the model) affects this trend, we see the performance by the number of reasoning steps provided by the GQA dataset evaluation, typically seen as the difficulty metric. Divided by this subcategory, ICL performs slightly better when the number of steps is larger (Table B.4). Together with LLaVA results, these results suggest that the performance boost may serve as a task difficulty indicator.

N Steps	N Samples	ZSL	ICL
1-5	12,153	59.7 ± 0.15	52.5 ± 0.31
6-9	65	83.5 ± 0.24	84.6 ± 0.27

Table B.4: Impact of multi-image ICL in GQA for InternVL 1b. N steps indicate the number of inference steps. The numbers with an error represent accuracy(%) in the corresponding setting. ICL boosted the performance when the number of steps was above six, implying that the ICL positively affects the performance when the task is challenging.

B.5 Other Considerations

B.5.1 Potential Risks

A hateful meme is a highly sensitive research topic. Therefore, all the hateful meme research involves risks and uncertainty to some extent. For example, the attackers may read a publication about a hateful meme detector to create a new meme that the detector may not be able to detect. More broadly, all LLM-related papers can be maliciously used when they are in the wrong hands (e.g., to improve an LLM trained on the dark web). To overcome these issues, an iterative update of the methodology with safety measures is a must.

B.5.2 Ethical Considerations

The hateful memes challenge dataset [18, 41] contains sensitive content. Therefore, we refrained from showing actual hateful memes so that this paper does not negatively impact any targeted group. We refer the users to the original publication for the considerations taken in dataset curation.

B.5.3 AI Assistant Usage

We used GitHub Copilot for efficient coding and ChatGPT for linguistic improvements.

B.5.4 License and Usage of Scientific Artifacts

We declare that all scientific artifacts used in this study do not prohibit the use of artifacts for academic research.

B.5.5 Documentation Of Artifacts

Experiment I uses the test split of six VQA datasets. GQA contains 10% of 22,669,678 questions over 113,018 images. TextVQA contains 5,734 text-image pairs. VizWiz contains 8,000 visual questions. VQAv2 contains 447,793 questions for 81,434 images. MMBench contains 1,784 questions. MM-Vet contains 218 questions.

Experiment II is performed on test-seen split of a hateful meme challenge dataset with 1,000 text-image pairs (510 benign samples and 490 hateful samples).

Correct ZSL response by LLaVA-1.5.

User:



What kind of beer is this?

Reference OCR token: NINK, [...],
ALE

Answer the question using a single
word or phrase.

Assistant:
ale

Incorrect ICL response by LLaVA-1.5.

User:



Provide a one-sentence caption for
the provided image.

Reference OCR token: ND-
CRAFTED, [...], PINE, BELT,
PALE, ALE, [...]

Assistant:
a aluminum can of southern star
pine belt ale

User:



What kind of beer is this?

Reference OCR token: NINK, [...],
ALE

Answer the question using a single
word or phrase.

Assistant:
pine belt

Figure B.2: A representative LLaVA-1.5’s response for which OoD ICL impacts the performance *negatively* from the TextVQA dataset. (Left) LLaVA-1.5 answered the **correct** object tag (**ale**). (Right) LLaVA-1.5 answered *incorrectly*, potentially caused by the wrong label (*pine* and *belt*) presented in the ICL example.

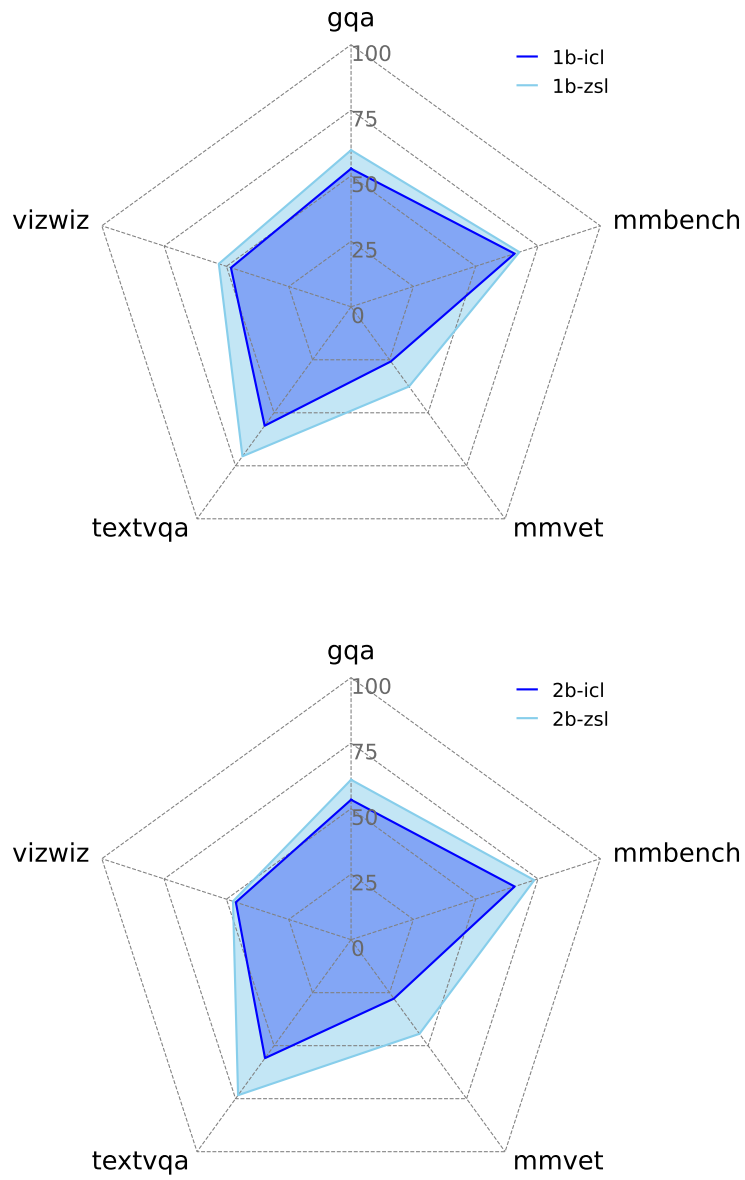


Figure B.3: Performance summary of InternVL. MM OoD dropped the performance for all the datasets, potentially reflecting that the baseline performance is moderate to high for all the datasets.

References

- [1] Niklas Muennighoff. Vilio: State-of-the-art Visio-Linguistic Models applied to Hateful Memes. *arXiv preprint*, December 2020.
- [2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *The Thirty-seventh Annual Conference on Neural Information Processing Systems*, New Orleans, LA, USA, December 2023.
- [3] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-Attention Attribution: Interpreting Information Interactions Inside Transformer. In *The 35th AAAI Conference on Artificial Intelligence*. AAAI Press, February 2021.
- [4] Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. On Explaining Multimodal Hateful Meme Detection Models. In *Proceedings of the ACM Web Conference 2022*, pages 3651–3655, Virtual Event, Lyon France, April 2022. ACM.
- [5] Jean-Marie John-Mathews. Where do biases in ML come from?, January 2022.
- [6] Pei Lee, Laks V.S. Lakshmanan, Mitul Tiwari, and Sam Shah. Modeling impression discounting in large-scale recommender systems. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1837–1846, New York New York USA, August 2014. ACM.
- [7] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

- [8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017.
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, pages 4171–4186, June 2019.
- [10] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training.
- [11] Radford, Alec, Wu, Jeff, Child, Rewon, Luan, David, Amodei, Dario, and Sutskever, Ilya. Language Models are Unsupervised Multitask Learners, 2019.
- [12] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., December 2020.
- [13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Thirty-First Annual Conference on Neural Information Processing Systems*, Long Beach, CA, USA, December 2017.
- [14] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A Survey of Large Language Models. *arXiv preprint*, June 2023.
- [15] Kilian Carolan, Laura Fennelly, and Alan F. Smeaton. A Review of Multi-Modal Large Language and Vision Models. *arXiv preprint*, March 2024.

- [16] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the Frontier of Vision-Language Models: A Survey of Current Methodologies and Future Directions. *arXiv preprint*, April 2024.
- [17] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias, May 2016.
- [18] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Thirty-Fourth Annual Conference on Neural Information Processing Systems*, Red Hook, NY, USA, December 2020.
- [19] Ziyang Xu, Keqin Peng, Liang Ding, Dacheng Tao, and Xiliang Lu. Take Care of Your Prompt Bias! Investigating and Mitigating Prompt Bias in Factual Knowledge Extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15552–15565, Torino, Italia, May 2024. ELRA and ICCL.
- [20] Ao Li, Zongfang Liu, Xinhua Li, Jinghui Zhang, Pengwei Wang, and Hu Wang. Modeling Variants of Prompts for Vision-Language Models. *arXiv preprint*, March 2025.
- [21] Daniel Reich and Tanja Schultz. Uncovering the Full Potential of Visual Grounding Methods in VQA. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4406–4419, Bangkok, Thailand, 2024. Association for Computational Linguistics.
- [22] Yanhan Sun, Jiangtao Qi, Zhenfang Zhu, Kefeng Li, Liang Zhao, and Lei Lv. Bias-guided margin loss for robust Visual Question Answering. *Information Processing & Management*, 62(2):103988, March 2025.
- [23] Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. Quantifying and Mitigating Unimodal Biases in Multimodal Large Language Models: A Causal Perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16449–16469, Miami, Florida, USA, 2024. Association for Computational Linguistics.
- [24] Boyue Wang, Xiaoqian Ju, Junbin Gao, Xiaoyan Li, Yongli Hu, and Baocai Yin. Counterfactual Dual-Bias VQA: A Multimodality Debias

- Learning for Robust Visual Question Answering. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13, 2025.
- [25] Trang Nguyen and Naoaki Okazaki. Causal Reasoning through Two Cognition Layers for Improving Generalization in Visual Question Answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9221–9236, Singapore, 2023. Association for Computational Linguistics.
 - [26] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing Vision-Language Models via Biased Prompts, May 2023.
 - [27] Congzhi Zhang, Linhai Zhang, Jialong Wu, Yulan He, and Deyu Zhou. Causal Prompting: Debiasing Large Language Model Prompting based on Front-Door Adjustment. In *The 39th Annual AAAI Conference on Artificial Intelligence (AAAI 25)*, Philadelphia, PA, USA, March 2025.
 - [28] Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. Inside-Out: Hidden Factual Knowledge in LLMs. *arXiv preprint*, March 2025.
 - [29] Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. MM-LLMs: Recent Advances in MultiModal Large Language Models, January 2024.
 - [30] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *arXiv:1707.07998 [Cs]*, Salt Lake City, UT, USA, June 2018.
 - [31] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems (NIPS 2019)*, Vancouver, BC, Canada, December 2019.
 - [32] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. VQA: Visual Question Answering. In *Fifteenth International Conference on Computer Vision (ICCV15)*, Santiago, Chile, December 2015.
 - [33] Hao Tan and Mohit Bansal. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and*

the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5100–5111, Hong Kong, China, November 2019. Association for Computational Linguistics.

- [34] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SIMVLM: SIMPLE VISUAL LANGUAGE MODEL PRE- TRAINING WITH WEAK SUPERVISION. In *The Tenth International Conference on Learning Representations (ICLR 2022)*, April 2022.
- [35] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A Visual Language Model for Few-Shot Learning. In *The Thirty-sixth Annual Conference on Neural Information Processing Systems*, volume 35, pages 23716–23736, New Orleans, LA, USA, November 2022. Curran Associates, Inc.
- [36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*, June 2023.
- [37] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, New Orleans, LA, USA, December 2023.
- [38] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A Jointly-Scaled Multilingual Language-Image Model.

In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, Kigali, Rwanda, May 2023.

- [39] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, The United States of America, July 2017. IEEE.
- [40] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language Is Not All You Need: Aligning Perception with Language Models. In *The Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS 2023)*, New Orleans, LA, USA, December 2023.
- [41] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Ron Zhu, Niklas Muennighoff, Riza Velicoglu, Jewgeni Rose, Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, Helen Yannakoudakis, Vlad Sandulescu, Umut Ozertem, Patrick Pantel, Lucia Specia, and Devi Parikh. The Hateful Memes Challenge: Competition Report. *Proceedings of Machine Learning Research*, 2021.
- [42] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-Task Vision and Language Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, page 10, Seattle, WA, USA, June 2020. IEEE.
- [43] Yuyang Chen and Feng Pan. Multimodal detection of hateful memes by applying a vision-language pre-training model. *PLOS ONE*, 17(9):e0274300, September 2022.
- [44] Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. Prompting for Multimodal Hateful Meme Classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pages 321–332, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [45] Ali Anaissi, Junaid Akram, Kunal Chaturvedi, and Ali Braytee. Detecting and Understanding Hateful Contents in Memes Through Captioning and Visual Question-Answering. *arXiv preprint*, April 2025.

- [46] Ahmed El-Sayed and Omar Nasr. AAST-NLP at Multimodal Hate Speech Event Detection 2024 : A Multimodal Approach for Classification of Text-Embedded Images Based on CLIP and BERT-Based Models. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 139–144, St. Julians, Malta, March 2024. Association for Computational Linguistics.
- [47] Piush Aggarwal, Jawar Mehrabian, and Weigang Huang. Text or Image? What is More Important in Cross-Domain Generalization Capabilities of Hate Meme Detection Models? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 104–117, St. Julian’s, Malta, March 2024. Association for Computational Linguistics.
- [48] Liam Kofi Bright, Daniel Malinsky, and Morgan Thompson. Causally Interpreting Intersectionality Theory. *Philosophy of Science*, 83(1):60–81, January 2016.
- [49] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179, September 2024.
- [50] Angelina Wang, Vikram V. Ramaswamy, and Olga Russakovsky. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 336–349, June 2022.
- [51] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual VQA: A Cause-Effect Look at Language Bias. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12695–12705, Nashville, TN, USA, June 2021. IEEE.
- [52] Qiwen Lu, Shengbo Chen, and Xiaoke Zhu. Collaborative Modality Fusion for Mitigating Language Bias in Visual Question Answering. *Journal of Imaging*, 10(3):56, February 2024.
- [53] Zefeng Zhang, Hengzhu Tang, Jiawei Sheng, Zhenyu Zhang, Yiming Ren, Zhenyang Li, Dawei Yin, Duohe Ma, and Tingwen Liu. Debiasing

- Multimodal Large Language Models via Noise-Aware Preference Optimization. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025 (CVPR 2025)*, Nashville, TN, USA, June 2025.
- [54] Yiting Qu, Xinlei He, Shannon Pierson, Michael Backes, Yang Zhang, and Savvas Zannettou. On the Evolution of (Hateful) Memes by Means of Multimodal Contrastive Learning. In *The 44th IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, May 2023.
 - [55] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and Mitigating Position Bias of Multi-image Vision-Language Models. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025 (CVPR 2025)*, Nashville, TN, USA, June 2025.
 - [56] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.
 - [57] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. INTERPRETABILITY IN THE WILD: A CIRCUIT FOR INDIRECT OBJECT IDENTIFICATION IN GPT-2 SMALL. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, Kigali, Rwanda, May 2023.
 - [58] nostalgebraist. Interpreting GPT: The logit lens, August 2020.
 - [59] Hanzhang Zhou, Zijian Feng, Zixiao Zhu, Junlang Qian, and Kezhi Mao. UniBias: Unveiling and Mitigating LLM Bias through Internal Attention and FFN Manipulation. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, Vancouver, BC, Canada, December 2024.
 - [60] Tony Z Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate Before Use: Improving Few-Shot Performance of Language Models. In *Proceedings of the 38 Th International Conference on Machine Learning (ICML 2021)*, volume 19. PMLR, July 2021.
 - [61] Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. Mitigating Label Biases for In-context Learning.

- [62] Zhixiong Han, Yaru Hao, Li Dong, Yutao Sun, and Furu Wei. PRO-TOTYPICAL CALIBRATION FOR FEW-SHOT LEARNING OF LANGUAGE MODELS. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, Kigali, Rwanda, May 2023. PMLR.
- [63] Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. Representation Engineering for Large-Language Models: Survey and Research Challenges. *arXiv preprint*, February 2025.
- [64] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24, Delft Netherlands, November 2023. ACM.
- [65] Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roei Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. Representation Surgery: Theory and Practice of Affine Steering. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, volume 235, Vienna, Austria, June 2025. PMLR.
- [66] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Venice, October 2017. IEEE.
- [67] Xingyu Fu, Ben Zhou, Sihao Chen, Mark Yatskar, and Dan Roth. Dynamic Clue Bottlenecks: Towards Interpretable-by-Design Visual Question Answering. *arXiv preprint*, April 2024.
- [68] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, December 2017.
- [69] Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States, 2022. Association for Computational Linguistics.

- [70] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online, 2021. Association for Computational Linguistics.
- [71] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2017.
- [72] Hoin Jung, Taeuk Jang, and Xiaoqian Wang. A Unified Debiasing Approach for Vision-Language Models across Modalities and Tasks. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, Vancouver, BC, Canada, December 2024.
- [73] Yandong Bi, Huajie Jiang, Yongli Hu, Yanfeng Sun, and Baocai Yin. Fair Attention Network for Robust Visual Question Answering. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9):7870–7881, September 2024.
- [74] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3319–3328. PMLR, August 2017.
- [75] Lloyd Stowell Shapley. A Value for n-person Games. In *Contributions to the Theory of Games*, volume II, pages 307–317. Princeton University Press, 1953.
- [76] Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, pages 512–515, Montreal, Canada, May 2017.
- [77] Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. MUTE: A Multimodal Dataset for Detecting Hateful Memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*. Association for Computational Linguistics, November 2022.
- [78] Minh Duc Bui. Multi3Hate: Multimodal, Multilingual, and Multicultural Hate Speech Detection with Vision–Language Models.

- [79] Drew A. Hudson and Christopher D. Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, June 2019. IEEE.
- [80] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, Salt Lake City, UT, USA, June 2018. IEEE.
- [81] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA Models That Can Read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8309–8318, Long Beach, CA, USA, June 2019. IEEE.
- [82] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is Your Multi-modal Model an All-around Player? In *WSDM ’23: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1128–1131, February 2023.
- [83] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities, October 2023.
- [84] Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online, 2021. Association for Computational Linguistics.
- [85] Wes Gurnee and Max Tegmark. Language Models Represent Space and Time. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria, May 2024.
- [86] Nathan Godey. On the Scaling Laws of Geographical Representation in Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and*

Evaluation (LREC-COLING 2024), pages 12416–12422, Torino, Italia, May 2024. ELRA and ICCL.

- [87] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv preprint*, January 2020.
- [88] Keyon Vafa, Justin Y. Chen, Jon Kleinberg, Sendhil Mullainathan, and Ashesh Rambachan. Evaluating the World Model Implicit in a Generative Model. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*, Vancouver, BC, Canada, December 2024.
- [89] Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. What Makes Chain-of-Thought Prompting Effective? A Counterfactual Study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1448–1535, Singapore, 2023. Association for Computational Linguistics.
- [90] SE Robertson, S Walker, MM Beaulieu, M Gatford, and A Payne. Okapi at TREC-4. In *The Fourth Text REtrieval Conference (TREC-4)*, page 73, January 1996.
- [91] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, July 2021.
- [92] Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. A Survey on Natural Language Counterfactual Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4798–4818, Miami, Florida, USA, 2024. Association for Computational Linguistics.
- [93] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An Explanation of In-context Learning as Implicit Bayesian Inference. In *The Tenth International Conference on Learning Representations*. arXiv, April 2022.
- [94] Ruifeng Ren and Yong Liu. In-context Learning with Transformer Is Really Equivalent to a Contrastive Learning Pattern, October 2023.

- [95] Matanel Oren, Michael Hassid, Yossi Adi, and Roy Schwartz. Transformers are Multi-State RNNs, January 2024.
- [96] Johannes von Oswald, Eyvind Niklasson, Randazzo, Ettore, Sacramento, Jo\~{a}o, Mordvintsev, Alexander, Zhmoginov, Andrey, and Vladymyrov, Max. Transformers Learn In-Context by Gradient Descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 1464, page 24, Honolulu, HI, USA, July 2023. JMLR.org.
- [97] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers. In *Findings of the Association for Computational Linguistics*, pages 4005–4019. Association for Computational Linguistics, July 2023.
- [98] Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. In-Context Learning of Large Language Models Explained as Kernel Regression, May 2023.
- [99] Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large Language Models Are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. In *The Thirty-seventh Annual Conference on Neural Information Processing Systems*, New Orleans, LA, USA, December 2023.
- [100] Yingcong Li, M. Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as Algorithms: Generalization and Stability in In-context Learning, February 2023.
- [101] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context Learning and Induction Heads, March 2022.
- [102] Hakaze Cho, Mariko Kato, Yoshihiro Sakai, and Naoya Inoue. RE-VISITING IN-CONTEXT LEARNING INFERENCE CIRCUIT IN LARGE LANGUAGE MODELS. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, Singapore, Singapore, April 2025.

- [103] Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. FUNCTION VECTORS IN LARGE LANGUAGE MODELS. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria, May 2024.
- [104] Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. Parallel Structures in Pre-training Data Yield In-Context Learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8582–8592, Bangkok, Thailand, 2024. Association for Computational Linguistics.
- [105] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *5th International Conference on Learning Representations (ICLR 2017)*, Toulon, France, April 2017.
- [106] Sebastian Farquhar and Yarin Gal. What ‘Out-of-distribution’ Is and Is Not. In *MLSafety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, New Orleans, LA, USA, November 2022.
- [107] Alan Ramponi and Barbara Plank. Neural Unsupervised Domain Adaptation in NLP—A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics.
- [108] Jingzhao Zhang, Aditya Krishna Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, and Suvrit Sra. COPING WITH LABEL SHIFT VIA DISTRIBUTIONALLY ROBUST OPTIMISATION. In *The Ninth International Conference on Learning Representations (ICLR 2021)*, May 2021.
- [109] Ruihan Wu, Chuan Guo, Yi Su, and Kilian Q Weinberger. Online Adaptation to Label Distribution Shift. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, December 2021.
- [110] Yi Dai, Hao Lang, Kaisheng Zeng, Fei Huang, and Yongbin Li. Exploring Large Language Models for Multi-Modal Out-of-Distribution Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5292–5305, Singapore, 2023. Association for Computational Linguistics.

- [111] Rena Gao, Xuotong Wu, Siwen Luo, Caren Han, and Feng Liu. ‘No’ Matters: Out-of-Distribution Detection in Multimodality Long Dialogue. *arXiv preprint*, October 2024.
- [112] Hao Lang, Yinhe Zheng, Binyuan Hui, Fei Huang, and Yongbin Li. Out-of-Domain Intent Detection Considering Multi-Turn Dialogue Contexts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12539–12552, Torino, Italia, May 2024. ELRA and ICCL.
- [113] Jiasheng Ye, Yawen Ouyang, Zhen Wu, and Xinyu Dai. Out-of-Distribution Generalization Challenge in Dialog State Tracking. In *Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, New Orleans, LA, USA, November 2022.
- [114] Xiangxi Shi and Stefan Lee. Benchmarking Out-of-Distribution Detection in Visual Question Answering. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5473–5483, Waikoloa, HI, USA, January 2024. IEEE.
- [115] Hugo Touvron, Louis Martin, and Kevin Stone. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*, July 2023.
- [116] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning, October 2023.
- [117] Henrik Singmann and David Kellen. An Introduction to Mixed Models for Experimental Psychology. In Daniel Spieler and Eric Schumacher, editors, *New Methods in Cognitive Psychology*, pages 4–31. Routledge, 1 edition, October 2019.
- [118] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *The 3rd International Conference for Learning Representations (ICLR 2015)*, San Diego, CA, USA, 2015. arXiv.
- [119] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Venice, Italy, October 2017.
- [120] Anton Razzhigaev, Matvey Mikhalechuk, Elizaveta Goncharova, Nikolai Gerasimenko, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov.

- Your Transformer is Secretly Linear. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5376–5384, Bangkok, Thailand, 2024. Association for Computational Linguistics.
- [121] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive Representation Learning: A Framework and Review. *IEEE Access*, 8:193907–193934, 2020.
 - [122] Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *The Thirty-seventh Annual Conference on Neural Information Processing Systems*, New Orleans, LA, USA, December 2023.
 - [123] Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing Model Behavior with Path Patching. *arXiv preprint*, May 2023.
 - [124] Xander Davies, Lauro Langosco, and David Krueger. Unifying Grokking and Double Descent. In *MLSafety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, New Orleans, LA, USA, November 2022.
 - [125] Leonard Bereska and Efstratios Gavves. Mechanistic Interpretability for AI Safety A Review. *Transactions on Machine Learning Research*, September 2024.
 - [126] DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu

- Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yuduan Wang, Yuheng Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhiniu Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui Gu, Zilin Li, and Ziwei Xie. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. *arXiv preprint*, June 2024.
- [127] Sumio Watanabe. A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research*, 14(1):31, January 2013.
- [128] Lucius Bushnaq, Jake Mendel, Stefan Heimersheim, Dan Braun, Nicholas Goldowsky-Dill, Kaarel Hänni, Cindy Wu, and Marius Hobbhahn. Using Degeneracy in the Loss Landscape for Mechanistic Interpretability. In *ICML 2024 Workshop on Mechanistic Interpretability*, Vienna, Austria, July 2024.
- [129] Vanessa Kosoy. Introduction To The Infra-Bayesianism Sequence, August 2020.
- [130] Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. Do LLMs Overcome Shortcut Learning? An Evaluation of Shortcut Challenges in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, page 13, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [131] Aaqib Syed, Can Rager, and Arthur Conmy. Attribution Patching Outperforms Automated Circuit Discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416, Miami, Florida, US, 2024. Association for Computational Linguistics.

- [132] Yuval Reif and Roy Schwartz. Beyond Performance: Quantifying and Mitigating Label Bias in LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6784–6798, Mexico City, Mexico, 2024. Association for Computational Linguistics.
- [133] Ilya Loshchilov and Frank Hutter. DECOUPLED WEIGHT DECAY REGULARIZATION. In *The Seventh International Conference on Learning Representations*, New Orleans, LA, USA, May 2019.
- [134] Shinichi Nakagawa and Holger Schielzeth. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142, February 2013.
- [135] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, June 2024. arXiv.

Publications

- [1] Yosuke Miyanishi and Minh Le Nguyen. 2024. Causal Intersectionality and Dual Form of Gradient Descent for Multimodal Analysis: A Case Study on Hateful Memes. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 2901–2916, Torino, Italia. ELRA and ICCL.
- [2] Yosuke Miyanishi and Minh Le Nguyen. 2025. Interpreting In-Context Learning for Semantics-Statistics Disentanglement via Out-of-Distribution Benchmark. In Submission.

