

**Modeling Visual Rhetorics for Persuasive Media through Self-supervised
Learning**

by

Meiqi Guo

Master of Science, Ecole Centrale Paris, 2018

Submitted to the Graduate Faculty of
the Department of Computer Science in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH
DEPARTMENT OF COMPUTER SCIENCE

This dissertation was presented

by

Meiqi Guo

It was defended on

June 28th 2023

and approved by

Dr. Rebecca Hwa, Department of Computer Science

Dr. Adriana Kovashka, Department of Computer Science

Dr. Diane Litman, Department of Computer Science

Dr. Daqing He, Department of Informatics and Networked Systems

Copyright © by Meiqi Guo
2023

Modeling Visual Rhetorics for Persuasive Media through Self-supervised Learning

Meiqi Guo, PhD

University of Pittsburgh, 2023

This dissertation addresses the challenging task of modeling and interpreting visual rhetorics in persuasive media using computational models. The focus is on self-supervised learning methods that leverage general data without specific annotations related to persuasion.

The research begins by modeling three fundamental modes of persuasion (ethos, pathos, logos) in multimodal media, incorporating both text and images. Traditional visual recognition models struggle to predict the applied persuasion modes in images beyond their literal content. Self-supervised learning methods prove to be more effective in modeling these modes. The detection of persuasive atypicality in ad images and the interpretation of symbolism are explored as common visual rhetorics for capturing viewers' attention and creating lasting impressions. The hypothesis that atypicality detection relies on contextual compatibility and understanding common-sense spatial relations of objects is validated through the development of self-supervised attention-based techniques. To assess the feasibility of automatically interpreting symbolism, an evaluative framework is developed. It compares the performance of language models and multi-modality models pretrained on large-scale web data. Furthermore, a re-ranking strategy is introduced to mitigate pre-training bias and significantly enhance model performance, bringing it on par with human performance in certain cases.

Overall, this dissertation presents a range of techniques that enable computational intelligence to detect, understand, and explain the underlying messages in rhetorical media. These methods leverage self-supervised learning and process large volumes of data, providing unprecedented depth and insight into the analysis of persuasive visual communication.

Keywords: Visual Rhetoric, Persuasion Mode, Persuasive Atypicality, Symbolism, Persuasive Media, Social Media, Advertisement Understanding, Self-supervised Learning.

Table of Contents

Preface	xii
1.0 Introduction	1
1.1 Motivation	1
1.2 Research Statement and Hypotheses	4
1.3 Thesis Overview	5
1.3.1 Modeling Modes of Persuasion	5
1.3.2 Detecting Atypicality	6
1.3.3 Interpreting Symbolism	6
1.4 Contributions	7
2.0 Background	10
2.1 Overview of Persuasions and Rhetorics	10
2.1.1 History and Evaluation of Rhetorical Theories	10
2.1.2 Visual Rhetorics	11
2.2 Computational Background for Modeling Rhetorics	14
2.2.1 Computational Approaches	15
2.2.2 Persuasion Dataset	17
2.3 Self-supervised Learning	19
2.3.1 Pre-training Objectives	19
2.3.2 Model Architectures	21
2.3.3 CLIP and Large Language Models	24
3.0 Modeling Modes of Persuasion	28
3.1 Introduction	28
3.2 Dataset Collection	29
3.2.1 Annotation Schemes	30
3.2.2 Annotation Strategies	33
3.2.3 Corpus Construction and Analysis	36

3.3	Political Ideology Bias in Topic-relevant Tweets	38
3.4	PersuaCLIP: Image Representation Learning from Tweet Text Supervision	40
3.4.1	Architecture of PersuaCLIP	41
3.4.2	Pre-training Objectives of PersuaCLIP	42
3.5	Experiments	43
3.5.1	Pre-training PersuaCLIP	44
3.5.2	Evaluation of PersuaCLIP with ImageArg	45
3.5.3	Limitations	48
3.6	Chapter Summary	49
4.0	Detecting Atypicality	50
4.1	Introduction	50
4.2	Our Approaches	52
4.2.1	Masked Region Reconstruction	52
4.2.2	Relative-Spatial Transformer	54
4.3	Experiments	56
4.3.1	Setup	56
4.3.2	Unsupervised Persuasive Atypicality Detection	60
4.3.3	Labelling Requirement for the Detection	63
4.3.4	Visual versus Semantic Compatibility	64
4.3.5	Limitations	66
4.4	Chapter Summary	67
5.0	Interpreting Symbolism	68
5.1	Introduction	68
5.2	Analysis Probe Construction	69
5.2.1	Symbolism Data Sources	70
5.2.2	Probing Methods	73
5.2.3	Analytical Tools	74
5.3	Re-ranking Approach for Bias Mitigation	75
5.4	Experiments	76
5.4.1	Setup	76

5.4.2	Model Performance on Decoding Symbolism	78
5.4.3	Effectiveness of Debiasing	79
5.4.4	Fine-grained Performance with Analytical Tools	82
5.4.5	Performance in Atypical Images	86
5.4.6	Limitations	87
5.5	Chapter Summary	88
6.0	Conclusion	89
6.1	Summary	89
6.2	Future Work	91
Appendix A. Annotation Instruction for ImageArg		93
A.1	Stance	93
A.1.1	Stance: Gun Control	93
A.1.2	Stance: Immigration	95
A.1.3	Stance: Abortion	97
A.2	Persuasiveness Level and Image Content	98
A.3	Persuasion Mode	104
Appendix B. Annotation Instruction for SymbA		108
Bibliography		109

List of Tables

Table 1: Inter-agreement for the first pilot annotation on <i>gun control</i> topic. . . .	34
Table 2: Inter-agreement after adjusting the annotation strategies on <i>gun control</i> topic.	34
Table 3: Inter-agreement rate of each annotation task on the topic <i>immigration</i> and <i>abortion</i>	36
Table 4: Image-text matching performance on tweet corpus.	45
Table 5: Prediction performance on modes of persuasion with image as input. . .	46
Table 6: Prediction performance on modes of persuasion with both image and text as input.	47
Table 7: The annotated description of atypical objects for each category.	58
Table 8: Performance of unsupervised models on the Ads dataset.	61
Table 9: Ablation study of layer number of encoder and relative positional feature.	62
Table 10: Experimental results on the Ads dataset.	64
Table 11: Comparison of Faster R-CNN RoI visual feature and predicted class label.	65
Table 12: Object detector performance on recognizing atypical objects.	66
Table 13: Signifier types of conventional literary symbolism.	71
Table 14: Human evaluation for decoding symbolism.	73
Table 15: Relationship type distribution in the set of advertising symbolism. . . .	75
Table 16: Model performance for decoding symbolism.	78
Table 17: Model performance on each signifier group of conventional symbolism. .	79
Table 18: Pearson correlation scores.	80
Table 19: Measuring the effectiveness of the re-ranking approach.	81
Table 20: Accuracy on the multi-choice task.	81
Table 21: Model performance on different PMI scores.	83
Table 22: The PMI score for each relationship type.	84
Table 23: Model performance on relationship types when using different prompts.	85

Table 24: Performance on atypical versus non-atypical advertising images. 86
Table 25: Summary of content types for each key purpose employed in the images. 104

List of Figures

Figure 1: Advertising images and social media posts that employ rhetorical devices.	3
Figure 2: Examples of atypical advertising images.	12
Figure 3: Examples of advertising images employed rhetorical devices.	13
Figure 4: Scheme of Attention.	23
Figure 5: Scheme of contrastive pre-training for CLIP.	25
Figure 6: The overview of our annotation pipeline.	31
Figure 7: Examples of image content types in tweets.	32
Figure 8: Examples of persuasion modes in tweet.	33
Figure 9: Examples of disagreement between annotators.	35
Figure 10: Distributions of stance, image persuasiveness, image content type, and image persuasion mode.	37
Figure 11: Distributions of image persuasiveness, content type and persuasion mode regarding stances.	37
Figure 12: Distributions of image content type regarding persuasion modes.	37
Figure 13: Example of a tweet containing statistical charts.	41
Figure 14: Examples where object interactions and their spatial relative position are important for atypicality detection.	51
Figure 15: Model overview.	53
Figure 16: Atypical object transformations in the Ads dataset.	57
Figure 17: Detection results of selected images from the Ads dataset.	62
Figure 18: A situated symbolism sample.	72
Figure 19: Knowledge difficulty distribution in the symbolism sets.	83
Figure 20: Case study for comparing predictions from RoBERTa and CLIP.	86
Figure 21: Example of stance annotation on gun control.	94
Figure 22: Example of stance annotation on immigration.	96
Figure 23: Example of stance annotation on abortion.	97

Figure 24:Example of a text only tweet.	98
Figure 25:Example of a tweet accompanying an image.	98
Figure 26:Example of tweets with statistics image and a non-statistics image.	99
Figure 27:Example of tweets with testimony image and a non-testimony image.	100
Figure 28:Example of tweets with anecdote image and a non-anecdote image.	100
Figure 29:Example of tweets with slogan image and a non-slogan image.	101
Figure 30:Example of tweets with scene photo image and a symbolic photo image.	102
Figure 31:Another example of tweets with scene photo image and a symbolic photo image.	103
Figure 32:Example of tweets with logos image and non-logos image.	105
Figure 33:Example of tweets with pathos images.	106
Figure 34:Example of tweets with ethos image and non-ethos image.	106

Preface

Completing this thesis marks the culmination of a transformative journey through the fascinating realm of natural language processing and multimodal analysis. As I reflect on this challenging yet rewarding endeavor, I find myself filled with immense gratitude and owe heartfelt appreciation to those who have supported and guided me throughout this academic odyssey.

First and foremost, I extend my deepest gratitude to my esteemed advisor, Prof. Rebecca Hwa, whose unwavering support and expertise in the field of natural language processing have been invaluable. Her critical thinking and insightful feedback have consistently enriched my research and paper writing, fostering a deeper understanding of the subject matter. Dr. Hwa's ability to ask probing questions has been instrumental in guiding my thought process, leading to innovative solutions and new research directions. I am grateful for the freedom and encouragement that she has granted me, allowing me to explore diverse research avenues and pursue my academic interests passionately. Her unwavering support has been the backbone of my Ph.D. journey, providing me with the necessary motivation to overcome challenges and reach new heights in my academic career. Moreover, I must express my sincere appreciation for Dr. Hwa's compassionate care. Her understanding of the challenges faced by students and her willingness to accommodate my needs by enabling remote work, have been instrumental in maintaining a healthy work-life balance during this transformative period.

Additionally, I extend my gratitude to the members of my committee, Prof. Adriana Kovashka, Prof. Diane Litman and Prof. Daqing He, for their valuable insights and feedback, which have played a crucial role in shaping the trajectory of this research. Their guidance has been invaluable in refining the thesis and elevating its scholarly merit. In addition to my advisor, I am grateful for the support and collaboration of several esteemed professors, Prof. Adriana Kovashka, Prof. Diane Litman, Prof. Malihe Alikhani, Prof. Siva Reddy, Prof. Yuru Lin, Prof. Wenting Chung, and Prof. Milos Hauskrecht, with whom I had the privilege to work closely during my Ph.D. journey. Their expertise, mentorship, and insightful discussions have been invaluable in shaping my research and expanding my horizons in the field of

natural language processing and machine learning. Their dedication to academic excellence and commitment to advancing research have inspired me to strive for greater heights in my academic pursuits.

I would like to acknowledge the support and camaraderie of my fellow researchers and lab mates, who have been an indispensable part of my academic journey. The collaborative environment at our institution has been a source of inspiration, fostering a stimulating intellectual atmosphere that nurtured my growth as a researcher. Their diverse perspectives and expertise have enriched my understanding of various topics and inspired new ideas for my research. Collaborating with them on projects and papers has been an enlightening experience, and their constructive feedback has helped me refine and strengthen my work.

I would also like to extend my appreciation to my mentors and colleagues during my internships at Google. Working at Google was a refreshing and rewarding break from academia, as it provided me with the opportunity to apply my research skills in real-world projects and collaborate with some of the brightest minds in the industry. Their expertise and insights into the industry have been instrumental in shaping my professional growth. They challenged me to think critically, approach problems from different angles, and encouraged me to push the boundaries of what I thought was possible.

Lastly, I express my heartfelt thanks to my family and friends for their unwavering belief in me and their constant encouragement throughout this journey. Their love and support have been my pillars of strength, allowing me to persist and thrive in the face of challenges. My thesis would not be complete without acknowledging the support and unconditional love that my beloved have given me. I cannot overstate my profound appreciation for my feline companion, Paris. Throughout the years of late-night research and intense study sessions, Paris has been a constant source of comfort and companionship. His playful presence and soothing purrs have helped me de-stress and stay grounded during challenging times.

As I conclude this preface, I am humbled by the experiences and knowledge I have gained throughout my Ph.D. years. I hope that this thesis can contribute, even in a small way, to the ever-expanding body of knowledge in modeling visual rhetorics. May it inspire future scholars to embark on their own intellectual adventures and contribute meaningfully to the world of research.

1.0 Introduction

1.1 Motivation

There is a growing interest in the automated understanding of persuasive media, such as commercial advertisements [50, 143, 109], political campaign broadcast [72, 111, 31], viewpoint-spreading tweets [80, 110, 141], etc. Rhetorics, being a potent tool for influencing and persuading, holds significant sway over the persuasive nature of such media [113, 76, 20]. To effectively deduce the intended message being conveyed, it becomes crucial for AI systems to model rhetorics in persuasive media.

Previous research efforts in modeling rhetorics have predominantly focused on media in textual or acoustic modality [142, 138, 137]. However, there is a dearth of exploration in the realm of visual modality and multi-modality, which incorporates both images and text. While language is widely recognized as the most effective medium for persuasive communication [17], visual rhetorics can significantly enhance persuasiveness. For instance, commercial posters often promote products by combining captivating imagery with catchy slogans. Recent studies have shown that social media posts with images garner higher popularity than those without [71], and images shared on social platforms significantly contribute to shaping political character development [82]. These applications and findings underscore the importance of developing computational models capable of detecting and interpreting visual rhetorics employed in persuasive multi-modal media.

Persuasion, in theory, involves one party attempting to influence another party to alter their opinion (believing or disbelieving something) or behavior (doing or not doing something) [8, 85]. The exploration of rhetoric and persuasion traces back to Aristotle, who proposed three fundamental modes of persuasion - ethos, pathos, and logos. These modes serve as common strategies in rhetoric, classifying how a speaker or writer appeals to their audience [25]. Recent work in computational models of rhetorics has been built upon this foundation. Previous research has demonstrated the application of these modes in analyzing the social impacts of discourses within social/environmental reports [46], as well as assessing

the persuasiveness of scientific texts [96] or student essays [16]. However, limited attention has been given to leveraging these modes to analyze persuasive images. We argue that despite the different modalities of media, the means of persuasion should remain consistent. Exploring the feasibility of modeling the modes of persuasion in persuasive images represents the initial step towards understanding and modeling visual rhetorics.

While the fundamental modes of persuasion provide the overarching strategies in rhetoric, a diverse array of rhetorical devices is essential for achieving the persuasive strategies discussed above. Rhetorical devices serve as powerful tools that enhance the persuasiveness, impact, and memorability of speeches, writing, or visual media when employed effectively with an audience. Observably, various forms of media exhibit distinct inclinations towards rhetorical devices. For instance, alliteration finds particular utility in speeches, whereas visual media tends to encapsulate a wealth of vivid information, adhering to the notion that “A picture is worth a thousand words”. However, this abundance of visual content also introduces a greater degree of ambiguity and uncertainty in interpretation. Consequently, identifying specific rhetorical devices employed in images or videos proves considerably more challenging than in written texts or speeches.¹ Therefore, when analyzing visual content, it becomes more meaningful to investigate rhetorical devices that are not only commonly employed but also possess broader applicability, such as *atypicality* and *symbolism*. Atypical portrayals intentionally crafted for persuasive purposes may involve imaginative object transformations, utilizing devices like metaphor, hyperbole, irony, and unique visual techniques. These techniques can include presenting an object with the texture of another object (*e.g.*, an apple with the texture of a kiwi) or distorting an object into a different shape (*e.g.*, a twisted skeleton). Similarly, symbolic associations of objects can be metaphorical (*e.g.*, a beer depicted as fresh as an apple), analogical (*e.g.*, a tomato resembling the phone company Apple’s logo, evoking thoughts of the fruit apple as a symbol of temptation), or ironic (*e.g.*, a gun pointed at the person holding it, symbolizing self-harm). Atypical and symbolic images are particularly effective in appealing to the persuasion strategies of pathos, logos, and ethos. For example, consider Figure 1(e), where the image creates a powerful emotional

¹For instance, in written texts, distinguishing between metaphors and similes is relatively straightforward as similes often adhere to recognizable patterns such as “as...as...” or “like”. However, these explicit patterns cannot be easily applied to visual media.

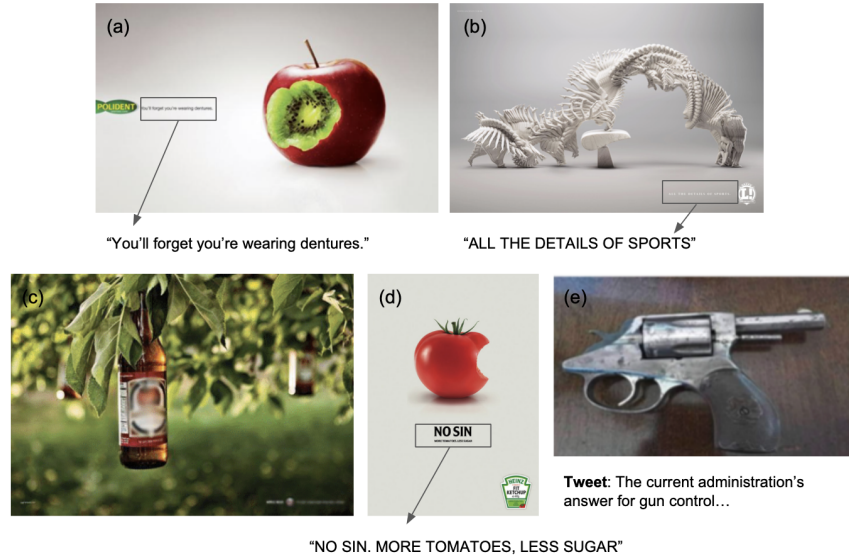


Figure 1: Advertising images [143] and social media posts [74] that employ rhetorical devices.

impact by symbolizing self-harm or danger. This evokes strong emotional responses, such as fear, concern, or empathy, depending on the context. Therefore, the ability to automatically detect atypicality and interpret symbolism plays a crucial role in developing computational intelligence capable of inferring the implied messages conveyed through rhetorical devices in persuasive media.

However, modeling visual rhetorics poses significant challenges for AI systems. Firstly, it necessitates a substantial amount of common-sense knowledge and higher-level reasoning beyond a literal interpretation of images [1]. For example, to classify whether an image evokes pathos in the persuasion mode, the intelligent system must understand the visual content that commonly elicits emotional responses in humans. If the rhetorical objective is achieved through atypical portrayals, knowledge of typical object interactions becomes crucial. Similarly, when objects are associated with symbolic purposes, the model needs to decode the symbolic references to infer implied messages. Secondly, scaling up the annotation of visual rhetoric is a daunting task. Models trained with only hundreds or thousands of annotated labels struggle to achieve sufficient performance. We argue that self-supervised learning on

large-scale data can offer a solution to these challenges. A carefully designed learning objective can empower models to capture extensive and demanding knowledge. Large language models have shown the ability to acquire factual and common-sense knowledge through pre-training [26, 52, 93]. The success of language models has also inspired the development of computer vision or multi-modal systems that learn representations from self-supervised signals [30, 75]. However, the applicability of this technique to modeling rhetoric, which requires understanding authors’ intentions and non-literal messages conveyed in the media, remains unclear. In this thesis, our goal is to address this question by developing computational models for visual rhetorics through self-supervised learning. By leveraging large-scale data and well-designed learning objectives, we aim to enable models to understand and interpret the complexities of rhetorical devices in visual media.

1.2 Research Statement and Hypotheses

By harnessing general data without persuasion-related labels, self-supervised learning methods can be developed to benefit the detection and interpretation of the visual rhetorics utilized in the persuasive media. In this thesis, we aim to test the following hypotheses:

H1. The modes of persuasion, originally used for analyzing speaking or textual media, can be adapted to effectively analyze persuasive images. The textual content accompanying the images serves as a valuable self-supervised signal for learning knowledge that aids in classifying the persuasion modes exhibited in the images.

H2. Atypical images can be detected by modeling contextual compatibility through self-supervised learning methodologies. Specifically, the interactions between objects and their spatial relative positions within the image play a significant role in the detection process.

H3. Language models and multi-modality models, trained through self-supervised learning on large-scale data, have acquired substantial knowledge of symbolism. This acquired knowledge can be leveraged to interpret the symbolism utilized in persuasive images.

1.3 Thesis Overview

This thesis is comprised of three distinct components, all of which share a common objective: modeling visual rhetorics for persuasive media using self-supervised learning. Each component explores a different aspect of rhetorics. The first part examines the three fundamental modes of persuasion in multi-modal media, such as Tweets that incorporate both text and images. Our research delves into understanding how these modes manifest in the combined textual and visual elements of persuasive messages. The second and third parts of the thesis concentrate on specific visual rhetorical devices. In the second part, we propose a novel model architecture and an efficient training objective to detect persuasive atypicality in advertising images. In the third part, we investigate the feasibility of automatically interpreting symbolism in persuasive media. To achieve this, we construct an evaluative framework that enables the analysis and understanding of symbolic references. Our aim is to unravel the implied messages conveyed through rhetorics. A brief overview of this thesis is presented below.

1.3.1 Modeling Modes of Persuasion

In Chapter 3, we delve into our research on modeling the modes of persuasion (ethos, pathos, logos) employed in persuasive images. To begin, we curate a comprehensive multi-modal dataset that includes annotations of image persuasiveness in tweets. Our investigation confirms that the three fundamental modes of persuasion can be effectively adapted for the analysis of persuasive images. One of the key challenges we encounter is the representation of images, which proves to be a bottleneck in the modeling process. However, we discover that the accompanying text in tweets contains valuable clues that can greatly contribute to predicting the persuasion modes utilized in the images. Leveraging this insight, we proceed to train a self-supervised model on a large-scale dataset of multi-modal tweets, utilizing the tweet text as the supervision signal. By employing the text as a form of guidance during the self-supervised learning process, we are able to enhance our model’s understanding of the persuasion modes exhibited within the images.

1.3.2 Detecting Atypicality

In Chapter 4, our objective is to detect persuasive atypicality in advertising images. This form of atypicality serves a specific purpose of conveying meaning and often involves metaphorical object transformations. Understanding the common-sense spatial relations between objects plays a crucial role in identifying persuasive atypicality. Based on this premise, we propose an approach that utilizes contextual compatibility as a self-supervised signal. To enable precise modeling of the spatial relations between objects, we introduce a novel method for computing attention weights. This method facilitates a more accurate representation of the interactions between objects in the image. Through extensive experimentation on a visual advertising dataset, we provide empirical evidence showcasing the effectiveness of our approach, particularly in detecting atypicality transformations that involve spatial interactions between objects.

1.3.3 Interpreting Symbolism

In Chapter 5, we delve into the interpretation of symbolism employed in advertising images. One of the major challenges in this task is that understanding the symbolic relationship between an object and a concept often relies on contextual and cultural knowledge, involving a complex chain of implicit reasoning. To address whether such specific symbolic knowledge can be captured through large-scale self-supervised learning, we introduce an evaluative framework. Within this framework, we compare the performance of language models and multi-modality models in interpreting different types of symbolism, utilizing various metrics for analysis. The results uncover the detrimental impact of biases inherent in pre-trained corpora. Additionally, we demonstrate the effectiveness of a simple re-ranking strategy in mitigating bias and significantly improving model performance, reaching a level comparable to human performance in certain cases.

1.4 Contributions

Visual rhetorics play a crucial role in making persuasive media more impactful, memorable, and persuasive to the audience. However, accurately detecting and interpreting these rhetorical devices in visual media presents unique challenges. This thesis presents a comprehensive investigation into the modeling of visual rhetorics for persuasive media through self-supervised learning, encompassing three distinct parts of research. With a unifying goal of understanding and capturing the persuasive elements present in visual media, each part delves into a different facet of rhetorics.

- To model modes of persuasion [74]:
 - We introduce a new multi-modal dataset called *ImageArg*, which serves as a valuable resource for annotating image persuasiveness in tweets. The creation of this dataset involves the development of novel strategies and schemes, ensuring the accuracy and reliability of the annotated persuasiveness labels.
 - We study the mutual influences between the modes of persuasion and other factors such as persuasiveness, visual content, and political ideology. By exploring the intricate relationships between these elements, we gain a deeper understanding of how persuasion operates in the context of multi-modal media.
 - We demonstrate the effectiveness of pre-training models by leveraging the textual content accompanying the images as a self-supervised signal. By capitalizing on the wealth of information present in the text, we show that pre-trained models can capture substantial knowledge related to persuasion.
- To detect persuasive atypicality [35]:
 - We propose a novel approach that leverages the implicit knowledge of contextual compatibility to detect persuasive atypicality, by reconstructing masked regions within the images as a self-supervised objective. Additionally, we explore the feasibility of modeling semantic compatibility as an alternative approach to detect atypicality. By comparing the performance of both approaches, we gain valuable insights into the effectiveness of different modeling strategies.

- We address the challenge of interpreting object-object spatial interactions within the images. To overcome this challenge, we introduce a new method for computing attention weights between key-query regions within our transformer-based models. This method enables a more precise modeling of the spatial relationships between objects, enhancing the accuracy of detecting persuasive atypicality.
- We conduct extensive experiments to evaluate the effectiveness of our proposed approaches. Furthermore, we analyze and interpret the results to gain insights into the impact of different types of persuasive atypicality on the overall detection performance.
- To interpret symbolism [36]:
 - We develop an evaluative framework called *SymbA*, which serves as a comprehensive resource for comparing and assessing the ability of different models, particularly language models, in decoding symbolism. We curate two sets of evaluative data that emphasize different aspects of symbolic relationships. Furthermore, we provide fine-grained categorizations of the evaluative data, enabling a deeper understanding of the characteristics of symbolic relationships that pose the greatest challenges to the models.
 - We address the issue of bias present in language models, which often favor commonly signified concepts. We propose quantification methods to measure the extent of bias in language models and develop techniques to mitigate this bias. Through empirical experiments, we demonstrate the effectiveness of our debiasing method, showcasing improvements in the performance of advanced language models.
 - We explore the capabilities of a pre-trained multi-modality model in capturing significant knowledge of symbolism. By leveraging the power of this state-of-the-art model, which has been trained on large-scale data encompassing both images and text, we demonstrate its effectiveness in interpreting symbolism employed in persuasive images.
 - We conduct a performance comparison of pre-trained models for decoding symbolism in advertising images, distinguishing between atypical and non-atypical instances.

Overall, our contributions in the field of modeling visual rhetorics provide novel methodologies, resources, tools, frameworks, and insights for researchers and practitioners. The findings and methodologies presented in this research have the potential to have a profound impact on various domains, including advertising, marketing, social media analysis, and content creation. By effectively detecting and interpreting the persuasive elements in visual media, computational models can assist advertisers and marketers in creating more impactful and persuasive campaigns. Moreover, understanding the rhetorical devices employed in persuasive images can help researchers and analysts gain deeper insights into the strategies used to influence public opinion, shape narratives, and convey messages through visual media. The development of accurate and robust computational models for visual rhetorics also opens up possibilities for enhancing content recommendation systems, sentiment analysis, and understanding the societal impact of persuasive media. Ultimately, this research aims to bridge the gap between human cognition and computational analysis, leading to advancements in our understanding of visual communication and paving the way for more sophisticated and intelligent systems in the domain of persuasive media analysis.

2.0 Background

In this chapter, we first present an overview of persuasion and rhetorics. We then review the computational background for processing, analyzing, and interpreting persuasion and rhetorics. Lastly, we review the literature of self-supervised learning that we use as the main technical method in this dissertation.

2.1 Overview of Persuasions and Rhetorics

Persuasion is a fundamental aspect of human communication that aims to achieve two major goals: influencing beliefs [8] and shaping behaviors [85]. Throughout history, individuals and organizations have employed various strategies to persuade others, including the use of argumentation structures and rhetorical devices. This section provides an overview of persuasion, its goals, and the role of rhetorical devices in shaping persuasive communication.

2.1.1 History and Evaluation of Rhetorical Theories

The study of persuasion can be traced back to ancient Greece, where scholars such as Aristotle, Plato, and Cicero laid the foundation for rhetorical theory. In their works, they identified key elements of effective persuasion, including ethos, pathos, and logos. Ethos appeals to credibility and authority, pathos evokes emotions and empathy, while logos relies on logical reasoning and evidence. These modes serve as the foundation for understanding persuasive communication and are often intertwined with specific rhetorical devices.

Over time, scholars and thinkers from various disciplines have contributed to the development of rhetorical theories. From the modern theories of Kenneth Burke, who emphasized the importance of identification and shared values [14], to the cognitive approaches of Richard E. Petty and John T. Cacioppo, who explored the cognitive processes underlying persuasion [95, 94], these theories have expanded our understanding of how persuasive messages are

constructed and received.

While traditional rhetorical theories primarily focused on spoken and written communication, the emergence of visual media has necessitated the exploration of visual rhetorics. Visual rhetorics examine the persuasive power of images, symbols, and visual representations. Scholars in this field analyze how visual elements convey meaning, evoke emotions, and influence audience perception. Understanding visual rhetorics is crucial in today's visual-centric society, where images play a significant role in advertising, social media, and political communication.

2.1.2 Visual Rhetorics

Visual rhetoric refers to the use of visual elements, such as images, graphics, and design, to communicate persuasive messages. Though the media modality is different, visual rhetoric operates on the same principles of persuasion as verbal and written rhetoric, aiming to influence beliefs and behaviors. Visual elements have the power to capture attention, evoke emotions, and convey complex ideas in a concise and impactful manner, thus can also be analyzed by Aristotle's modes of persuasion. For instance, a tweet featuring a statistical chart correlating gun fatalities with gun ownership to support the argument for stricter gun control exemplifies the use of *logos*, appealing to logic and evidence. A classic example of *ethos* in advertising is the ubiquitous "9 out of 10 dentists recommend this toothpaste" type of commercial, which relies on the credibility and authority of dental professionals to establish trust and persuade consumers. An impactful example of *pathos* in visual rhetoric can be observed in the use of photos depicting a hungry Syrian child with innocent eyes that seem to convey a heartfelt plea for help. Such images have the power to evoke strong emotions and compassion in the audience, thereby influencing their receptiveness to more lenient immigration policies for refugees. Occasionally, an image may incorporate multiple modes of persuasion, blending *ethos*, *pathos*, and *logos* harmoniously to enhance the persuasive impact. These three modes of persuasion are interconnected, working in synergy to bolster the strength of an argument. Collectively, they are often referred to as the rhetorical triangle, highlighting the interdependence and complementary nature of these persuasive strategies.

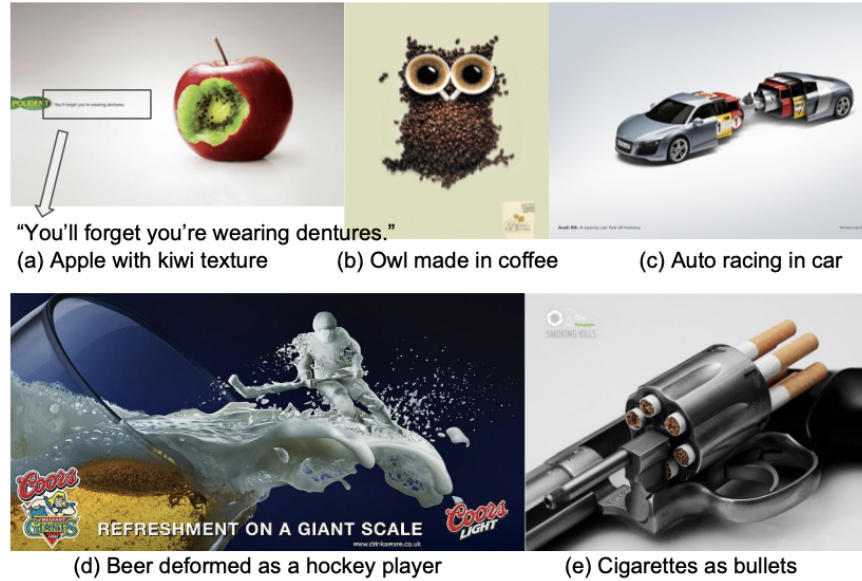


Figure 2: Examples of atypical advertising images.

Rhetorical devices are powerful tools employed in persuasive communication to make messages more impactful and memorable to the audience. These devices go beyond the general modes of persuasion and provide specific techniques for achieving persuasive goals. Rhetorical devices enable communicators to create compelling narratives, establish connections with the audience, and influence their attitudes and behaviors.

Creative and atypical portrayal of objects is a common and effective rhetoric for impressing audience with a persuasive purpose in visual media [47]. For example, an advertisement may use unusual imagery or storytelling techniques to stand out and leave a lasting impression on viewers. We show some atypical advertising images in Figure 2. Each of them employ different modes of persuasion. By drawing a kiwi inside an apple, the commercial ad sends the message that biting an apple is as easy as biting a kiwi with their dentures, by logos. Since the kiwi-apple object rarely appears in real life and requires some reasoning for figuring out the conveyed message, its persuasion influence on receivers lasts longer than directly saying the message in sentences. Another rhetorical technique that commonly exists in visual and multi-modal media is symbolism. Persuasive images are not merely analogues to visual



Figure 3: Examples of advertising images employed rhetorical devices.

perception but symbolic artifacts constructed from the conventions of cultural knowledge as well as common experiences [112]. The visual arguments are embodied in the visual arts. For example, the metaphorical object transformations for creating atypicality usually involve the use of symbolism. Taking examples in Figure 2, kiwi symbolizes soft, owl symbolizes invigorating, racing car symbolizes high-performance and speed, gun symbolizes danger, etc.

Additionally, these two powerful rhetorical devices, atypicality and symbolism, are closely intertwined with other rhetorical devices such as metaphor and hyperbole. Metaphor is a figure of speech that involves making a comparison between two seemingly unrelated things, highlighting their similarities to evoke a deeper understanding or create a vivid image. Like symbols, metaphors also replace some intended target concept with different ones; however, a metaphor emphasizes *some common property* it shares with the target concept. In contrast, a symbol serves as a *stand-in* for a more complex and abstract concept under certain context; it may not share any obvious property with the abstract concept, and it may not be associated with solely one concept [61]. Metaphor can be considered as a type of the generalized symbolism. Atypicality can be employed within a metaphor to create unexpected or unconventional comparisons, adding layers of meaning and engaging the audience in a thought-provoking manner. For instance, in Figure 3, the left image portrays an advertisement for a sports brand that aims to convey the concept of speed and agility. The image

features a running woman whose shadow intriguingly resembles the distinctive shape of a cheetah. This atypical depiction of a human’s shadow resembling a cheetah metaphorically suggests that the woman, when wearing the brand’s sportive equipment, runs as fast as a cheetah. By using an unexpected combination of elements, the image captures the viewer’s attention and prompts them to associate the desirable attributes of the cheetah with the sports brand. Hyperbole, on the other hand, is a rhetorical device that involves exaggerated statements or claims not meant to be taken literally. It is frequently used in atypical images. Taking the example of the right image in Figure 3, an advertisement for a GPS service depicts a person with an extraordinarily complex and atypical arm pointing towards a direction. This hyperbolic representation exaggerates the complexity to an extreme extent, emphasizing the idea that finding the road without GPS is exceptionally challenging. This hyperbolic representation creates a visually striking and attention-grabbing image that communicates the message of the product’s ability to provide an extraordinary and powerful experience. Furthermore, symbolism can be effectively incorporated within hyperbole to enhance its impact. By assigning symbolic meaning to certain elements or exaggerating their qualities, hyperbolic statements become more visually and emotionally compelling.

Understanding and analyzing rhetorical devices in persuasive communication is essential for several reasons. Firstly, it provides insights into the mechanisms behind successful persuasion, allowing communicators to craft more persuasive messages. Secondly, studying rhetorical devices contributes to the fields of communication studies, linguistics, and psychology, enhancing our understanding of human behavior and decision-making processes. Finally, in the era of digital media and information overload, the ability to recognize and evaluate rhetorical devices becomes crucial for media literacy and critical thinking.

2.2 Computational Background for Modeling Rhetorics

Developing computational approaches for the automated processing, analysis, and interpretation of persuasion holds immense potential across various domains. Firstly, such approaches can be employed to assess the quality of arguments or persuasion strategies.

For instance, they can be utilized to score and provide feedback on student essays, predict the effectiveness of advertising designs before their public release, and guide strategies for political campaigns. By analyzing the predicted persuasion modes in their campaign visuals, political teams can identify which persuasive techniques are resonating with their target audience. For example, they can determine whether appeals to emotion (Pathos) or logical reasoning (Logos) are more successful in conveying their messages. Armed with this knowledge, campaigns can refine their communication tactics to maximize their persuasive impact and engage voters more effectively. Secondly, modeling persuasion is crucial in the emerging field of argument search, where the aim is to locate persuasive materials related to a specific topic of interest from various modalities such as text, image, speech, and more. By utilizing computational approach to decode symbolism and detect atypicality in persuasive imagery, search engines and recommendation systems can identify relevant and persuasive materials related to specific topics of interest. For researchers, marketers, and content creators, this can streamline the process of finding compelling arguments and persuasive examples to support their work. Thirdly, in the realm of social media and online platforms, computational models for visual rhetorics can play a critical role in detecting and flagging instances of harmful and abusive persuasive content. By identifying atypicality and inferring symbolic meanings in images and text, automated content moderation systems can proactively detect harmful messages, such as racist or hateful content, before they spread widely. This can help in creating a safer online environment and mitigating the negative impact of persuasive techniques used for malicious purposes. In summary, the integration of computational models for visual rhetoric can enhance decision-making processes, improve persuasive interactions, and contribute to a more informed and responsible use of persuasion in different contexts.

2.2.1 Computational Approaches

In the field of argument mining and generation, several computational models have been developed to analyze argumentative structures in free text. These models examine elements such as conclusions, premises, inference schemes, and pro- and con-relations within textual arguments [91, 16, 34, 120, 121, 63]. However, these approaches are limited to the text

modality and cannot be extended to analyze persuasive multi-modal media. For example, when analyzing persuasive messages conveyed through images, these models consider the image as a whole without the ability to classify specific regions as premises, claims, or major claims. To address this limitation, our focus is on rhetorical devices that are employed in both textual and visual modalities, as it is essential to consider multiple modalities holistically when modeling persuasion [87].

While there has been some work on visual rhetorics that utilize facial expressions, bodily gestures, and scene context to analyze the communicative intents of images, these studies have predominantly focused on images of politicians [54, 49]. Therefore, there is a pressing need for computational investigations into a broader range of common rhetorical techniques utilized in various persuasive multi-modal media across different applications.

Regarding the modeling of fundamental modes of persuasion, prior work has mainly focused on language use. For instance, Higgins and Walker demonstrate how persuasion modes contribute to the social effects of discourses in social/environment reports [46]. Recent studies have examined the relationship between persuasiveness and persuasion modes in scientific texts [96] or student essays [16]. Hidey *et al.* analyze the order of premises within each mode present in online persuasive forums [45]. Some studies have concentrated on one or two specific modes, such as mining ethos in political debates [31] or classifying rhetorical questions into logos or pathos [38].

In terms of modeling specific rhetorical devices, there has been related work on recognizing metaphoric usages in natural language processing [83, 66], computer vision [122], and multi-modal media [117, 144]. Symbolism interpretation aligns with metaphor interpretation, aiming to establish connections between surface and target concepts [106, 116, 129, 56]. Previous approaches have explored connecting symbolism and metaphor through shared features or logical sequences, but such connections may not exist for symbolism. One prior work utilized an image encoder pretrained on ImageNet [27] to interpret symbolism in advertising images [50]. A subsequent study argued that standard image-based predictions alone are insufficient for symbolism prediction and demonstrated that additional fine-tuning of a language model for processing OCR-extracted text from ads can significantly improve performance [109]. The detection of persuasive atypicality is also an under-explored problem.

Existing work focuses on detecting atypical objects in real-world images, such as diverse defects [10] or out-of-context objects and scenes [23, 108], without a specific rhetorical objective.

2.2.2 Persuasion Dataset

To support the development of computational models, it is essential to have access to a persuasion dataset with labeled data. Numerous instances of people persuading others can be found in various online activities, providing opportunities for data collection. Previous research has explored different aspects of persuasion, including influence on beliefs or attitude change through online discussions [51] or stance change towards social topics [76]. Additionally, researchers have measured the strength of persuasion by examining human decisions or actions, such as lending loans [142], donating to charity [137], purchasing products [97], funding projects [79], and more. However, much of this work has primarily focused on language and speech as the communication modality [19, 138, 18].

In recent years, the study of argumentative relations across multiple modalities has garnered increasing attention. Researchers have made notable contributions in this area, such as Alikhani *et al.*, who annotated discourse relations between text and accompanying imagery in recipe instructions [3]. Kruk *et al.* investigated multi-modal document intent in Instagram posts [59], while Zhang *et al.* examined the implicit relationship between persuasive images and text [145]. These studies exemplify the interest in understanding how different modalities interact in persuasive communication. Within the domain of visual rhetorics, several multimodal datasets have been released to facilitate research on metaphor comprehension [144, 1]. These datasets provide valuable resources for studying the intricate relationship between language and visual elements, contributing to a deeper understanding of how metaphors are conveyed and interpreted in multimodal contexts. However, there is currently a lack of existing multimodal datasets that specifically analyze the modes of persuasion employed in images. While there are datasets available for studying persuasion and understanding metaphor in multimodal contexts, there is a gap when it comes to comprehensive analysis and annotation of the modes of persuasion in images. This highlights the

need for further research and the development of new datasets that focus on exploring the various modes of persuasion utilized in visual media.

Our thesis leverages an existing dataset released by Ye *et al.* [143], which is specifically designed for interpreting visual rhetorical devices in advertising. Advertising images are intentionally designed by experts to create associations in viewers' minds, often containing atypical and symbolically-associated objects. The Ads dataset comprises a substantial collection of advertising images, around 65k, providing ample data instances for self-supervised learning. A significant portion of these images are annotated with atypicality labels (around 4k) and symbolism annotations (around 14k). The atypicality annotation includes a binary label indicating whether the image is atypical or not. If it is atypical, the image is further classified into one or several atypicality categories, such as texture replacement, object with missing part, deformed object, and more. This level of granularity in classification provides valuable information for developing and analyzing computational models. Additionally, the dataset provides a sentence that describes the atypical objects present in the image. These detailed textual descriptions further enhance the dataset's utility for studying and modeling atypicality in visual media. The symbolism annotation involves drawing bounding boxes on the image to indicate the location of symbolic objects, along with their corresponding symbolic references. However, it is important to note that a textual description specifically focusing on the symbolic part of the image is not provided in the dataset. This limitation hampers the precise analysis of models' performance in interpreting symbolism. Having a textual modality to represent the symbolic images would be beneficial as it would enable the isolation of models' capability in interpreting symbolism from potential visual recognition issues that could indirectly impact their performance. This distinction is crucial because a model's performance in decoding symbolism may be hindered by inadequate object recognition. By separating the analysis of symbolism from visual recognition challenges, researchers can gain a clearer understanding of the models' specific capabilities and limitations in interpreting symbolism. Overall, this dataset serves as a valuable resource for investigating persuasion in visual media, particularly within the context of advertising. The inclusion of atypical and symbolically-associated objects in the dataset enables the further exploration of the persuasive impact of these rhetorical devices.

2.3 Self-supervised Learning

In recent years, self-supervised learning has emerged as a powerful paradigm in machine learning and artificial intelligence. Traditionally, supervised learning has been the dominant approach, where models are trained on labeled data provided by human annotators. However, acquiring large amounts of labeled data can be expensive, time-consuming, and sometimes infeasible due to the need for expert annotation. Self-supervised learning, on the other hand, aims to alleviate the reliance on labeled data by leveraging the inherent structure and information present in unlabeled data. It involves designing tasks or objectives that can be automatically generated from the data itself, allowing the model to learn useful representations without explicit human supervision. By exploiting the vast amounts of unlabeled data available, self-supervised learning enables the training of deep and complex models with millions or even billions of parameters.

Researchers have focused on addressing two key questions in the development of efficient self-supervised learning methods. First, they aim to define effective pre-training objectives that facilitate the learning of generalized and robust knowledge that can be applied to different tasks [28, 103, 21]. This involves designing tasks or objectives that encourage the model to learn meaningful representations and capture important features of the data. Second, researchers strive to develop models that can be efficiently trained using self-supervised data. This includes the design of architectures and learning algorithms that are well-suited for self-supervised learning settings. In this section, prior efforts related to these questions are discussed, highlighting the progress made in self-supervised learning. Furthermore, several notable pre-trained models that are utilized in this dissertation are introduced, showcasing their contributions to the field.

2.3.1 Pre-training Objectives

The core idea behind self-supervised learning is to formulate a pretext task that requires the model to make meaningful predictions or decisions about the data. These pretext tasks are carefully designed to encourage the model to capture essential features, structures, or

patterns in the data, which can then be transferred to downstream tasks of interest. The key advantage of self-supervised learning is that it enables models to learn rich and generalized representations from large-scale unlabeled data, which can be further fine-tuned on smaller labeled datasets for specific tasks.

In natural language processing (NLP), self-supervised learning has been successfully applied to language models. One common approach is to train language models to predict the next token in a sequence given the preceding context, known as auto-regressive language models. This allows the model to capture the semantic and syntactic properties of language, leading to impressive results in tasks such as text generation, sentiment analysis, and machine translation. Models such as GPT (Generative Pre-trained Transformer) [102, 103, 13] fall into this category. Another popular approach in NLP is the use of masked language models, exemplified by models like BERT (Bidirectional Encoder Representations from Transformers) and its variations [28, 73]. In this approach, a subset of tokens in the input is masked, and the model is trained to predict the masked tokens based on the surrounding context. This encourages the model to learn deeper semantic understanding and contextual dependencies, enabling it to excel in tasks such as text classification, named entity recognition, and question answering. Masked language models often outperform auto-regressive models when fine-tuned on downstream NLP tasks. However, they may underperform in text generation tasks due to the masking scheme and the assumption of independence between masked tokens [133].

Self-supervised learning has also made significant strides in computer vision. While supervised learning with labeled image datasets, such as ImageNet [27], has been widely used [135, 6], self-supervised learning offers an alternative by leveraging unlabeled image data. Unlike language, the raw signal in vision is continuous and high-dimensional. This has led to the exploration of various pretext tasks for learning visual representations, such as colorization [147, 62, 131], jigsaw puzzles [84, 29], inpainting [89], instance discrimination [140], recovering the input from corruptions [130], and contrastive learning between transformed images [42, 21]. By training models to solve these pretext tasks, they can learn powerful visual representations that can be applied to tasks like image classification, object detection, and semantic segmentation. Furthermore, self-supervised learning has expanded beyond in-

dividual modalities and has ventured into the realm of multi-modal learning [75, 124, 11, 81]. Given that multi-modal signals are often complementary in real-world applications [7, 145, 4], learning representations in a multi-modal setting has gained attention. By combining information from different modalities such as text, images, and audio, models can learn richer and more comprehensive representations. Common pre-training objectives in images and text include matching image-text pairs [127, 100], matching images with text and predicted object tags [70, 146], aligning words with visual regions [22] and more. This opens up new possibilities for applications that involve multiple modalities, such as image captioning and visual question answering.

While self-supervised learning has shown great promise in training deep neural networks without extensive labeled data, its effectiveness in understanding the persuasive aspects of visual rhetoric remains uncertain. Although pre-trained models have demonstrated their proficiency in comprehending the literal visual content of images for tasks such as classification and captioning, it remains to be seen whether they have acquired knowledge related to the persuasive elements employed in visual rhetoric. Understanding the communicative message conveyed by visual rhetoric requires a deeper level of analysis and interpretation beyond the recognition of objects and scenes. It involves capturing the intended symbolism, atypicality, and other rhetorical devices used to influence the audience’s perception and behavior. To address this gap, new approaches and methodologies need to be developed to enhance the capability of self-supervised learning methods in capturing and interpreting these persuasive elements.

2.3.2 Model Architectures

Self-supervised learning has gained significant attention in recent years as a powerful technique for training deep neural networks without relying on large amounts of labeled data. One of the key factors contributing to the success of self-supervised learning is the choice of model architectures that can effectively learn meaningful representations from unlabeled data. In this section, we discuss some of the popular model architectures used in self-supervised learning.

In the early days of self-supervised learning, traditional model architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and basic Neural Networks (NNs) played a crucial role in representing and learning from unlabeled data. CNNs have been the go-to architecture for computer vision tasks for over a decade, excelling at extracting visual features from raw pixel data [64, 43, 118]. By leveraging convolutional layers and pooling operations, CNNs capture spatial hierarchies and local patterns in images, enabling them to learn powerful representations. On the other hand, RNNs are commonly used for processing sequential data such as language [92], allowing them to capture temporal dependencies and context in sequential information. Word2Vec is another classic architecture used for learning language representation [78]. This architecture involves a two-layer neural network that operates on the context of words in a large corpus of text. It represents words as dense vectors in a continuous space, capturing semantic relationships between words based on their co-occurrence patterns in a large corpus of text. These model architectures have laid the groundwork for self-supervised learning by demonstrating the ability to learn meaningful representations from unlabeled data. While they still hold value and are applied in certain scenarios, they have been complemented and surpassed in recent years by more advanced models like transformers.

Transformers have emerged as a groundbreaking architecture in the field of self-supervised learning and have revolutionized natural language processing and computer vision tasks. Originally introduced by Vaswani *et al.* in the context of machine translation [128], transformers have proven to be highly effective in capturing long-range dependencies and contextual relationships in sequential data. The key innovation of transformers lies in their self-attention mechanism (shown in Figure 4), which allows the model to weigh the importance of different parts of the input sequence when making predictions. This attention mechanism enables transformers to capture global dependencies and handle variable-length inputs more efficiently compared to traditional recurrent neural networks (RNNs). Unlike RNNs, transformers process the entire input sequence in parallel, eliminating the sequential nature of computations and making them highly parallelizable and well-suited for distributed training on GPUs. This makes transformers well-suited for large-scale training with vast amounts of data.

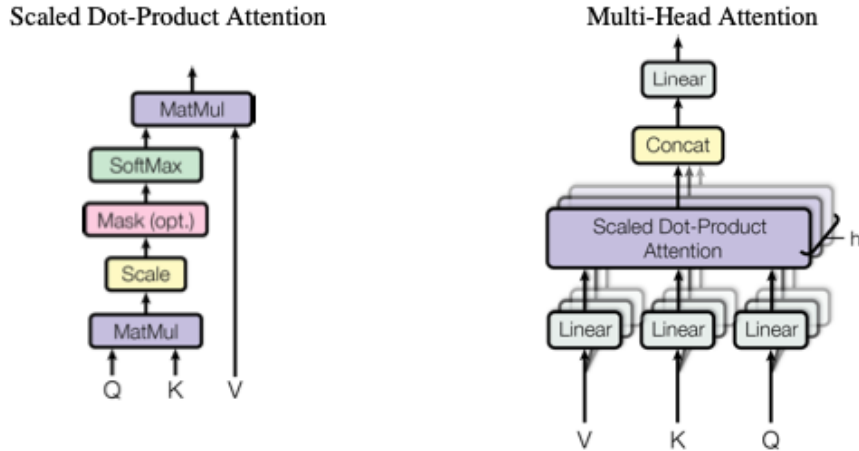


Figure 4: Scheme of Attention: (left) Scaled Dot-Product Attention; (right) Multi-Head Attention consists of several attention layers running in parallel [128].

In addition to their success in natural language processing [103, 28], transformers have also demonstrated remarkable performance in computer vision tasks [88, 15, 30] and vision-language tasks [127, 75, 65, 123, 148, 98, 68, 22]. Vision transformers (ViTs) [30] apply the transformer architecture to image data by dividing the input image into patches and treating them as sequential data. By leveraging self-attention mechanisms, ViTs can capture spatial dependencies and effectively model long-range interactions between image patches. This enables them to capture high-level semantic information and achieve competitive performance in image classification, object detection, and vision-language tasks.

However, it's important to note that using a sequence of image patches may overlook a crucial aspect of visual content interpretation: the spatial relationships between objects. While transformers can capture contextual dependencies within the patches, the spatial arrangement of objects plays a significant role in understanding visual scenes and the conveyed message, especially in persuasive images that employ visual rhetorics. To address this limitation, further research can be conducted to better model the spatial interactions between objects in visual data. This involves designing novel architectures or modifications to existing transformer-based models that explicitly incorporate spatial information and capture

the relationships between image patches.

2.3.3 CLIP and Large Language Models

We introduce several notable pre-trained models that are utilized in this dissertation.

CLIP (Contrastive Language-Image Pretraining) is a state-of-the-art model that has gained significant attention in the field of vision and language understanding [100]. It has demonstrated remarkable performance in various vision tasks such as optical character recognition), action recognition in videos, geo-localization, and many types of fine-grained object classification. Developed by Radford *et al.* at OpenAI, CLIP aims to bridge the gap between visual and textual domains by jointly training a neural network on a large corpus of image and text pairs.

At the core of CLIP is a contrastive learning framework, which enables the model to learn meaningful representations directly from the data. Unlike traditional methods that rely on explicit labels or annotations, CLIP learns by contrasting positive pairs (consisting of an image and its associated text) against negative pairs (combinations of different images and texts). The scheme is shown in Figure 5. By optimizing the model to differentiate between positive and negative pairs, CLIP learns to associate semantically related visual and textual elements. The training objective of CLIP is to maximize the agreement between image and text representations, while minimizing the agreement between mismatched pairs. This objective encourages the model to capture the underlying semantic similarities and correspondences between images and their textual descriptions. By leveraging this contrastive learning approach, CLIP can learn to understand the nuanced relationships and contextual information present in visual and textual data, which is necessary for understanding visual rhetorics.

In terms of model architecture, CLIP employs a transformer-based neural network, which allows for the modeling of complex relationships and dependencies in both images and text. In the case of images, CLIP divides them into patches and treats them as a sequence of data, similar to how sentences are treated as sequences of tokens in natural language processing. By leveraging this architecture, CLIP captures both local and global information within images,

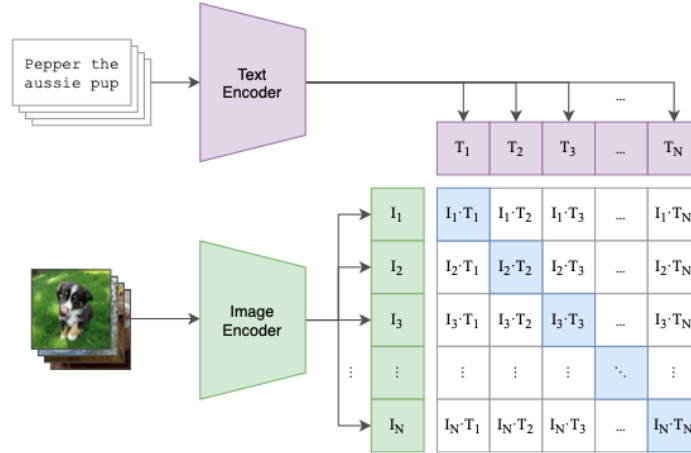


Figure 5: Scheme of contrastive pre-training for CLIP [100].

allowing it to understand objects, their spatial relationships, and the broader context in which they appear. This capability is essential for analyzing visual rhetorics, as it enables the model to discern the intended messages conveyed through symbolic objects, compositional choices, and other persuasive visual techniques.

Furthermore, CLIP benefits from its large-scale pre-training on diverse and extensive datasets, encompassing a wide range of visual concepts and linguistic patterns. This pre-training process enables the model to acquire a broad understanding of visual and textual semantics, including persuasive cues and rhetorical devices employed in images. This is crucial for understanding visual rhetorics, as persuasive images often exhibit unique and unconventional features that require the model to go beyond simple object recognition.

In summary, CLIP’s ability to capture complex relationships between images and text, its generalization capabilities, and its robust visual representations make it an ideal candidate for modeling visual rhetorics. Through its contrastive learning approach and transformer-based architecture, CLIP is expected to excel in understanding and interpreting the persuasive impact of visual elements, enabling deeper analysis of the communicative messages conveyed through visual rhetorics.

LLM (Large Language Models), such as BERT, RoBERTa, and the GPT series, have

revolutionized the field of natural language processing by learning vast amounts of knowledge from large-scale pre-training. These models excel at capturing and encoding complex linguistic patterns, semantic relationships, and contextual information, enabling them to achieve impressive performance on various language understanding tasks.

BERT leverages masked language modeling to capture rich linguistic knowledge, including syntactic structures, semantic meanings, and even subtle nuances [28]. RoBERTa (Robustly Optimized BERT Approach)[73] builds upon BERT’s success by further optimizing its training methodology. RoBERTa removes certain training objectives used in BERT, such as next sentence prediction, and focuses on longer pre-training times and larger amounts of training data. By extensively tuning hyperparameters and scaling up the training process, RoBERTa achieves improved generalization and outperforms BERT on various NLP benchmarks. The additional training enables RoBERTa to learn more nuanced linguistic knowledge, including domain-specific information and deeper semantic representations. GPT models are trained to predict the next word in a sequence given the preceding context. Through massive-scale pre-training on diverse and extensive text data, GPT models acquire a comprehensive understanding of language. They learn grammar, syntax, semantics, and even world knowledge, allowing them to generate coherent and contextually appropriate text. GPT-3, with its massive size of 175 billion parameters, has demonstrated unprecedented language generation capabilities, producing human-like text across various tasks and domains [13]. Specifically, large language models have the potential to interpret rhetorical devices due to their exposure to a wide range of texts from diverse sources. They are trained on vast corpora of text, encompassing different genres, styles, and topics. This exposure exposes them to a broad spectrum of rhetorical devices used in various contexts, such as literature, news articles, scientific papers, online forums, and social media.

The knowledge learned by these large language models is encoded in their parameters and can be transferred to a wide range of downstream tasks. A common approach is to fine-tune the pre-trained models on specific tasks. Additionally, prompting methods have been developed to elicit knowledge from language models by providing explicit textual prompts. GPT-3, for example, has demonstrated the ability to solve tasks when given a textual prompt with only a few examples, sometimes even achieving competitive results compared to prior

state-of-the-art fine-tuning approaches. These prompts typically involve masking certain information, such as filling in a missing word in a sentence (e.g., “Barack Obama was born in [MASK]”). The model can generate the missing information based on its learned knowledge and context [93, 44]. To improve the effectiveness of prompts, researchers have proposed various methods, including mining, paraphrasing, or learning from a training corpus to generate better prompts [52, 26, 115]. The use of explicit textual prompts allows for a more controlled and directed interaction with the language models, enabling targeted querying and elicitation of specific knowledge. These approaches provide a means to extract valuable information and insights from the models without extensive fine-tuning or explicit supervision. These techniques provide flexibility and efficiency in leveraging the knowledge learned by LLMs, opening up avenues for diverse applications and analysis.

In summary, large language models have the potential to possess learned knowledge for interpreting rhetorical devices due to their exposure to diverse textual data, contextual understanding, self-supervised learning, and transfer learning capabilities. Their ability to capture patterns, semantic relationships, and contextual cues equips them with the necessary foundation to comprehend and analyze the persuasive strategies employed through rhetorical devices. Leveraging this knowledge can aid in developing more sophisticated models and facilitate better analysis and interpretation of rhetorics.

3.0 Modeling Modes of Persuasion

3.1 Introduction

In this chapter, we focus on developing methods to automatically identify the mode of persuasion used in multi-modal media aimed at changing people’s beliefs. While previous research has primarily focused on analyzing argumentation in language-based texts such as scientific papers and student essays [96, 16], there is a significant opportunity to leverage other modalities, such as images, which can potentially enhance the persuasiveness of the argument. By investigating the applicability of the three modes of persuasion to analyze persuasive images, we aim to gain insights into the rhetorical strategies employed in multi-modal media.

To bridge this gap, we address the lack of exploration in image persuasiveness by developing computational models that predict the modes of persuasion employed in images. However, the absence of an annotated dataset poses a challenge, as discussed in Section 2.2.2. To overcome this, we create a new multi-modal dataset called *ImageArg*, specifically designed to annotate image persuasiveness in tweets, thereby extending the domain of persuasiveness mining into the multi-modal realm. In Section 3.2, we discuss the challenges encountered during the creation of this corpus, including the design choices made and the methodology employed to ensure its validity.

The analysis conducted using *ImageArg* reveals a strong correlation between human political ideology (i.e., stance towards a social topic) and the argumentative features present in their posted tweets. This finding emphasizes the need to address political ideology bias that may inadvertently be present in the annotated dataset. Unbalanced distributions of tweets favoring or opposing a stance can introduce bias into computational models trained on skewed data. Therefore, we explore the interactions between political ideology bias and topic-relevance classifiers, and propose a method to mitigate this bias in Section 3.3.

A key challenge encountered in our research lies in the representation of images, which proves to be a bottleneck in predicting the modes of persuasion. However, we make an impor-

tant discovery that the accompanying text in tweets contains valuable clues that significantly contribute to understanding the persuasion strategies employed in the images. Building upon this insight, we train a self-supervised model on a large-scale dataset of multi-modal tweets, utilizing the tweet text as a form of supervision signal (hypothesis **H1**). By leveraging the text as guidance during the self-supervised learning process, we enhance our model’s understanding of the persuasion modes manifested within the images, as discussed in Section 3.4.

3.2 Dataset Collection

We collect a multi-modal dataset, *ImageArg*, consisting of annotations of image persuasiveness in tweets. We choose tweets as the data source because they frequently contain both image and text and the image in a tweet greatly increases its global influence to audience [71]. Moreover, we retrieve tweets that are relevant to a social topic, *e.g.* gun control, because this kind of tweets usually try to convince an audience to support their social stance, thus are persuasive. However, there exist several challenges for annotating the dataset. First, the existing annotation schemes were previously developed to capture the persuasive strength of text arguments in essays [31, 132, 16]. We need to adjust and extend them in order to correctly analyze the persuasion of images in a multi-modal setting. Second, a novel taxonomy is required to annotate image content that explicitly identifies image functionalities in the argumentative aspect. Last but most importantly, prior work reveals that it is difficult to obtain a high agreement for annotating the modes of persuasion used in the text, especially for annotating pathos [38, 45]. It is expected to be even harder with images because communication in visual modality has a less capacity for effective reception of information than in textual modality [17]. We need to explore different annotation strategies for obtaining a high-quality annotation. We finally evaluate the inter-rater agreement on the annotations for demonstrating the feasibility to model the visual rhetoric with the modes of persuasion established by Aristotle.¹

¹The construction of this dataset was a collaborative work with Yue Dai and Zhexiong Liu [74]. I led the design of the annotation instruction and strategies. In addition, I was fully responsible for selecting qualified annotators through Amazon Mechanical Turk, creating annotation interface, launching annotation

3.2.1 Annotation Schemes

The annotations are based on a persuasion taxonomy we developed to explore image functionalities and the means of persuasion. We build a corpus of Twitter posts on a social topic (*e.g.*, gun control), then annotate each multi-modal post along four dimensions. The annotation pipeline is shown in Figure 6. First, we determine **(1)** the **stance** of the entire tweet because it contributes to extensive argumentation mining pipelines and potentially plays an essential role in the persuasiveness-related tasks [77]. Specifically, we assume one tweet holds a consistent stance in its text and image since the author would intend to deliver a consistent argument. For those tweets annotated with a positive or negative stance, we also annotate **(2)** the **persuasiveness score** of the tweet image and **(3)** the image **content type**. The content types identify image roles in the argumentative aspect by describing what kind of evidence is contained in the image for enhancing the persuasiveness of the whole tweet (*e.g.*, supportive data, authorized photos, *etc.*). Rather than looking at the image alone, the visual evidence is annotated by considering the argumentative relationship between the image and text. For those images annotated as persuasive, we identify its **(4)** **persuasion mode**, which indicates how the images persuade audiences (*e.g.*, by providing strong logic, touching audiences emotionally, *etc.*). The annotation scheme for each task is presented as following.

Stance. We use existing methods [80] to verify if the image holds a clear stance on a given topic. Specifically, given a tweet (including both text and images), we ask annotators to select among four stances: positive (*i.e.*, support), negative (*i.e.*, oppose), neutral, or irrelevant to the topic.

Image Persuasiveness. We adopt five levels of persuasiveness scores proposed in a prior work studying textual argumentation [16]:

- **(L0) No persuasiveness:** the annotated target fails to convince the audience at all.
- **(L1) Medium persuasiveness:** the annotated target partially convinces the audience.
- **(L2) Persuasive:** the annotated target is convincing to the audience.
- **(L3) High persuasiveness:** the annotated target is very convincing to the audience.

tasks, processing annotation results, computing inter-rater agreements, and conducting data analysis.

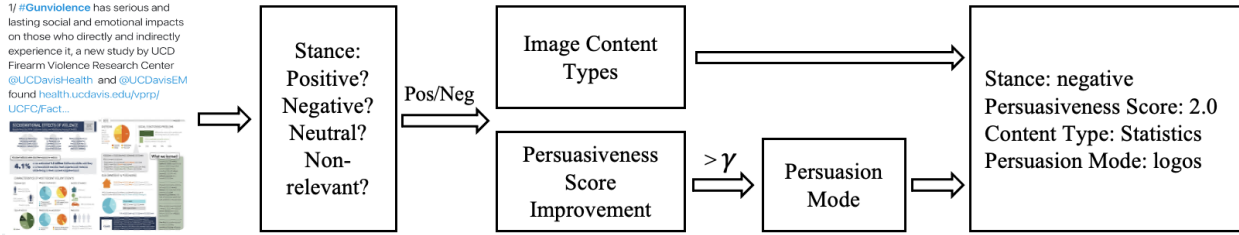


Figure 6: The overview of our annotation pipeline. Annotators start by annotating the argumentative stance of input tweets. Afterwards, tweets with either positive or negative stances are annotated for image content types and persuasiveness score improvement. The persuasion mode is further annotated if persuasiveness score improvement exceeds a given threshold γ ($\gamma = 0.5$ in this work).

- **(L4) Extreme persuasiveness:** the annotated target is compelling to the audience.

Instead of asking annotators how persuasive the image is, we propose to compute it by calculating the persuasiveness gain brought by the image. We first ask annotators to choose one of 5 persuasiveness levels based on pure text. Next, we ask annotators to give a second choice based on both text and image. Then we compute the difference between these two scores (image-text score minus text-only score) for measuring the persuasiveness of the image². For avoiding personal bias, the final score is computed as an average of three annotations. To interpret image persuasiveness, we use a threshold γ ($\gamma = 0.5$) that encodes the score into a binary label (i.e., persuasive or not).

Image Content. We leverage the definition of argumentative roles of evidence in news to categorize image content: Statistics, Testimony, and Anecdote [2]. However, these categories developed for text fail to capture all the image contents that frequently appear in tweets, for example, photographs. To this end, we propose three visual-dominant categories, Slogan, Scene photo and Symbolic photo. Six categories are defined for representing the content of images (examples are shown in Figure 7):

- **Statistics:** Images provide evidence by stating or quoting quantitative information, such

²We set the image persuasiveness score to 0 in the case of negative.

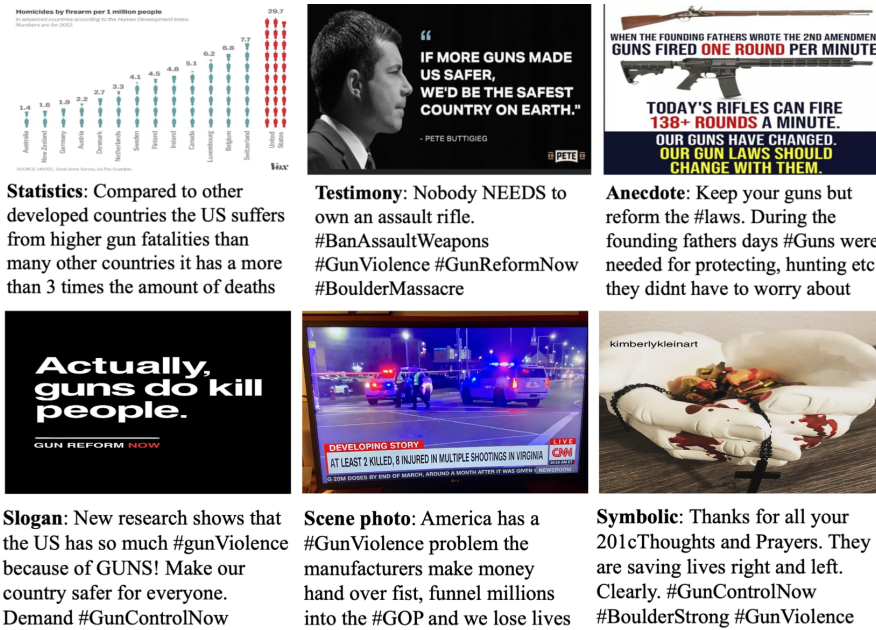


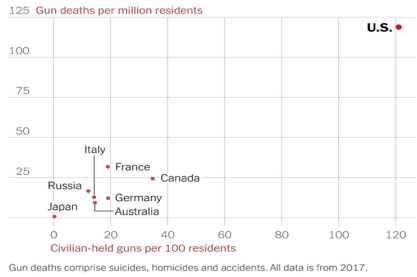
Figure 7: Examples of image content types in tweets: statistics, testimony, anecdote, slogan, scene photo, and symbolic.

as a chart or diagram showing data.

- **Testimony:** Images quote statements or conclusions from an authority, such as a piece of articles or claims from an official document.
- **Anecdote:** Images provide information based on the author’s personal experience, such as facts/personal stories.
- **Slogan:** Images embed pieces of advertising/slogan text.
- **Scene photo:** Images show a real scene or photograph.
- **Symbolic photo:** Images show a symbol/art that expresses the author’s viewpoints in a non-literal way.

Image Persuasion Modes. We follow the three fundamental modes of persuasion established by Aristotle [25]:

- **Logos:** The image appeals to logic and reasoning, which persuades audiences with reasoning from a fact/statistics/study case/scientific evidence. The Logos image in Figure 8



Logos: Gun deaths and gun ownership by population - by country. Hmm. Well, this doesn't take much effort to figure out why we've got such #GunViolence...

Pathos: A personal narrative - Dr. Sonya Lewis “We must reject helplessness and complacency and we must allow ourselves to feel the raw, sick”

Ethos: The US has 4.4 % of the world's population but 42% of gun violence. #guncontrol #gunviolence

Figure 8: Examples of persuasion modes in tweet: logos, pathos, and ethos.

provides a chart that shows the high gun deaths and the high gun ownership by the population of the US, which implies a logical relationship between gun death and gun ownership.

- **Pathos:** The image appeals to emotion, which evokes emotional impact that leads to higher persuasiveness. The Pathos image in Figure 8 provides art that shows the grieved “Uncle Sam” saying “no” with helplessness, which excites the desire to control guns.
- **Ethos:** The image appeals to ethics, which enhances credibility and trustworthiness. The Ethos image in Figure 8 takes a screenshot of the source of a report from New York Times, which increases credibility.

3.2.2 Annotation Strategies

Since we are the first to annotate the persuasiveness of tweet image in a multi-modal setting, there is no annotation material, such as coding manual, that we can directly use. Therefore, we develop annotation strategies based on several rounds of pilot annotations. Since crowd-sourcing annotators³ have a wide range of educational background, we employ

³Our annotations were conducted on the platform of Amazon Mechanical Turk.

Table 1: Inter-agreement for the first pilot annotation on *gun control* topic.

Task	Alpha	Count
Stance	64.5	87
Image content type	71.1	38
Image persuasion mode	19.9	38

Table 2: Inter-agreement after adjusting the annotation strategies on *gun control* topic.

Task	Alpha	Count
Stance	76.1	100
Image content type	64.6	72
Image persuasion mode - Logos	55.3	56
Image persuasion mode - Pathos	51.0	56
Image persuasion mode - Ethos	57.8	56

qualified workers who passed a well-designed qualitative test that evaluates the workers’ understanding on our annotation manual.

We start with annotating tweets about *gun control*. In the first-round, we ask annotators to make a single choice from multiple candidates for each annotation task. Table 1 shows the Krippendorff’s alpha score [57] for measuring the inter-rater agreement⁴. Based on the standard interpretation of alpha scores [60, 40], we conclude that annotations on stance and content type have a substantial inter-agreement; but the inter-agreement for annotating persuasion mode is slight. The results reveal that extending the logos-ethos-pathos scheme to the image modality has some difficulties. We observe that more than one persuasion mode may appear in a tweet image, or none of the persuasion mode may be applicable. To solve this problem, we modify the coding manual for annotating persuasion modes. We use three-label annotation that asks to choose yes/no for each persuasion mode, instead of using three-class annotation that asks to choose one persuasion mode from all the three. Moreover, the annotators are requested to justify their choice by giving a brief reasoning comment. As shown in Table 2, the inter-rater agreement on persuasion modes is greatly improved, from

⁴Note that the availability of annotation questions is based on the answer to the prior questions (Figure 6) therefore each task has different sample numbers.

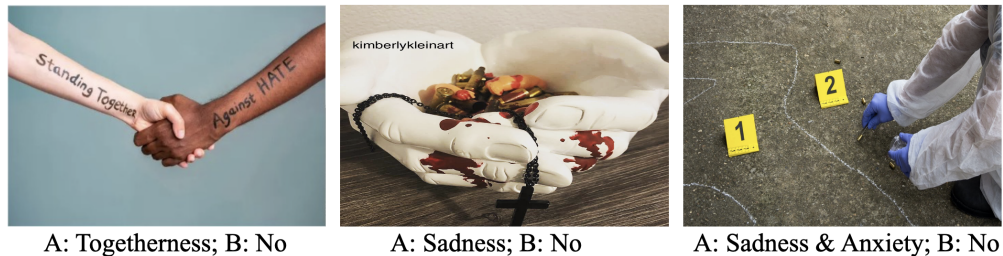


Figure 9: Examples of disagreement between annotators: annotator A annotates the above images as Pathos because these examples express emotions, while annotator B disagrees and marks as not Pathos.

19.9 to above 50. All the three modes achieve a moderate inter-rater agreement, while Pathos has the lowest agreement. This observation is consistent with previous work about annotating text [38, 45]. It is likely because annotators have different emotional reasoning (*i.e.*, some annotators are easily evoked by images while others are not). As the examples shown in Figure 9, one annotator recognized strong emotional impact (*e.g.*, togetherness, sadness, anxiety, *etc.*), while the other not.

We further perform pilot annotations for the topics of *immigration* and *abortion*, with the best annotation strategies that we developed for annotating *gun control*. We randomly choose 100 or 200 tweets respectively on *immigration* or *abortion* for the pilot study, and adjust the coding manual for annotating the stance by providing some topic-specific examples. The inter-rater agreement for both topics is shown in Table 3. We observe high inter-rater agreements on the stance annotation, which demonstrates the utility of our topic-specific coding manual. The agreement on the content type is generally good, however, *abortion* has relatively lower agreement than the other two topics. One main reason is that authors prefer using photos to support their arguments. Such photos lead to ambiguity between scene photos and symbolic photos. Moreover, we notice that the agreements on the persuasion modes are not satisfying. For *immigration*, Ethos has the lowest agreement. One explanation is that there are few authentic resources that provide credible and trustworthy arguments on this topic. For *abortion*, the agreement on all three persuasion modes are relatively

Table 3: Inter-agreement rate of each annotation task on the topic *immigration* and *abortion*.

Task	Immigration		Abortion	
	Alpha	Count	Alpha	Count
Stance	61.5	100	68.7	200
Content type	65.8	53	56.6	76
Logos	56.7	23	25.0	48
Pathos	46.0	23	37.5	48
Ethos	30.8	23	28.2	48

low, in particular, Logos surprisingly gets the lowest agreement. These studies indicate that the inter-rater agreement on annotating persuasion mode is topic-dependent, and the relationship between topics and persuasion modes needs further investigation. We thus collect annotations only on the *gun control* topic, and leave the other two topics for future work. The annotation instruction can be found in Appendix A.

3.2.3 Corpus Construction and Analysis

We collect raw tweets containing both image and text through TwitterAPI⁵. For ensuring that our retrieved tweets are relevant to a specific topic, we use the expert-selected keywords provided in a previous work [37]. We then retain tweets whose texts tend to be argumentative, with an argument confidence score larger than 0.9 by using ArgumentText Classify API⁶. This filtering process ensures our annotation data has high argumentation-confidence.

We annotate 1003 samples that hold a supporting or opposing stance towards *gun control*. The distribution of each annotation task is reported in Figure 10: (a) the distribution of stance is almost balanced; (b) surprisingly, only a quarter of images are persuasive⁷; (c) the most frequent content types are vision-dominant (*i.e.* Symbolic photo and Scene photo); text-dominant content (*i.e.* Anecdote, Slogan, Testimony) also occupy a significant propor-

⁵<https://developer.twitter.com/en/docs/twitter-api>

⁶<https://api.argumentsearch.com>

⁷The threshold γ is set to 0.5 in our annotations since the persuasiveness score is an average of three annotators, thus γ greater than 0.5 suggests that there is at least two annotators annotating images persuasiveness with L1 or higher (≥ 1) scores or at least one annotator annotating L2 or higher scores (≥ 2).

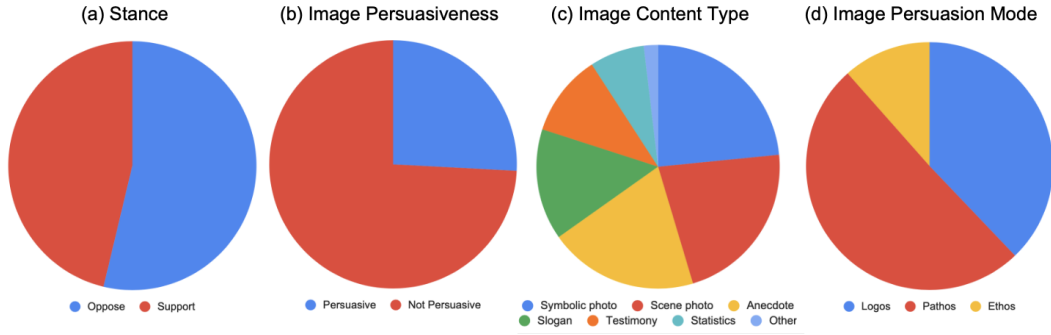


Figure 10: Distributions of (a) stance, (b) image persuasiveness, (c) image content type, and (d) image persuasion mode in our corpus.

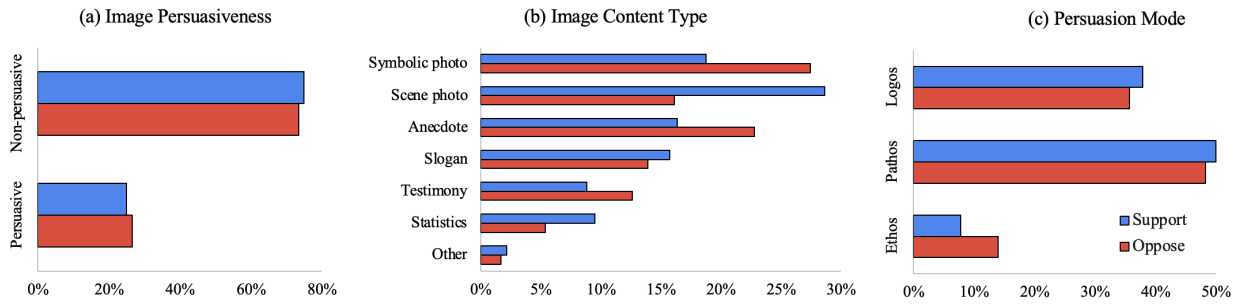


Figure 11: Distributions of (a) image persuasiveness, (b) content type and (c) persuasion mode regarding stances (support in blue and oppose in red) in our corpus.

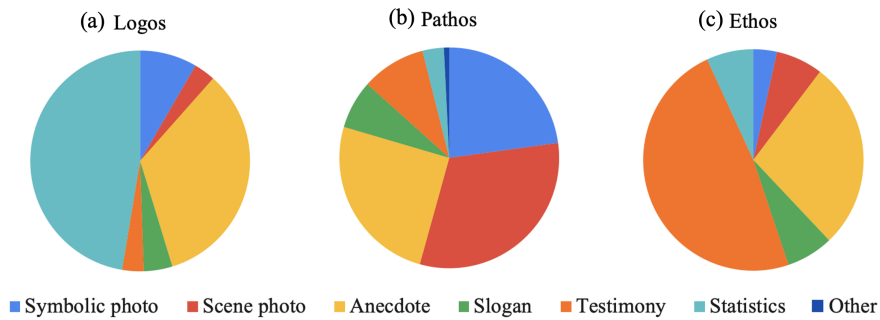


Figure 12: Distributions of image content type in different persuasion mode (a) Logos, (b) Pathos, and (c) Ethos in our corpus.

tion; in contract, the data-dominant content, Statistics, is the least frequent type,⁸ (d) nearly half of the tweet images persuade an audience by the emotional appeal while Ethos has the least of usage.

Next, we show how the stance relates to the image persuasiveness, content type, and persuasion mode in Figure 11: (a) supporting and opposing stance are almost evenly distributed in persuasive or non-persuasive images, which suggests that the persuasiveness of images is independent with the stance of arguments; (b) tweets holding an opposing stance uses significantly more images in the type of Symbolic photos, Anecdote, and Testimony; while tweets holding a supporting stance prefers attaching an image in the content of Scene photos or Statistics; (c) images supporting *gun control* applies more Logos and Pathos but less Ethos as their rhetorical strategy than those in the opposing stance.

To further study the relevance between image content type and persuasion mode, we report their correlated distributions in Figure 12: (a) most Logos images contain Statistics and Anecdote evidence, which meets the intuition that the logical reasoning can usually be clarified by introducing anecdotes or justified by providing supportive statistics; (b) the majority of Pathos images are Scene and Symbolic photos, which is as expected since images generally promote emotional impression by presenting visual information; (c) nearly half of the Ethos images contain Testimony because statements from authorities can enhance trustworthiness. These correlations imply mutual influences between different annotation dimensions and raise the possibility to use them as clues for modeling persuasion modes.

3.3 Political Ideology Bias in Topic-relevant Tweets

We address the issue of political ideology bias in the collection of topic-relevant tweets. We recognize the importance of constructing a dataset with a balanced distribution of political stances to ensure unbiased modeling of rhetorical strategies in persuasive media. In our study on collecting ImageArg, we employ a specific methodology to ensure minimal

⁸The sample is annotated as “Other” if the annotator thinks none of the content types can accurately describe the image. There is only 1.89% of Others, which demonstrates that our defined scheme for content types has a good coverage.

bias in the annotated data. We start by using a set of expert-selected keywords to retrieve topic-relevant tweets. Subsequently, we incorporate human annotation as the initial task in our pipeline, where we ask human annotators to annotate the stance of the tweets. This approach allows us to closely monitor the political ideology distribution and ensure that the annotated data exhibits minimal bias.

While human annotation guarantees minimal bias in our small-sized sample, it is not feasible for large-scale studies. To handle larger datasets, automatic topic detection models are typically used for extracting relevant text about the topic of interest from a vast data source. However, these models may inadvertently introduce or propagate biases, leading to skewed data collections and potentially incorrect conclusions. Therefore, it is crucial to develop accurate and unbiased topic detection models to collect reliable data.

To understand the impact of biased keywords on downstream training and retrieval, we conduct empirical analyses using three commonly used language models: GloVe [90], ELMo [92], and BERT [28]. Our findings indicate that BERT, among the three models, is more prone to propagating bias and experiencing a drop in retrieval quality when trained on biased data.

To mitigate this bias, we propose an approach that adapts Domain-Adversarial Training [33] for the three off-the-shelf models. That is, we want a classifier that is oblivious to an instance’s political ideology, yet still performs the main task of judging the instance’s relevance to the topic. Experimental results demonstrate that our approach effectively reduces unintended bias without significantly sacrificing retrieval accuracy. In fact, the debiased BERT model shows a slight improvement in retrieval accuracy.

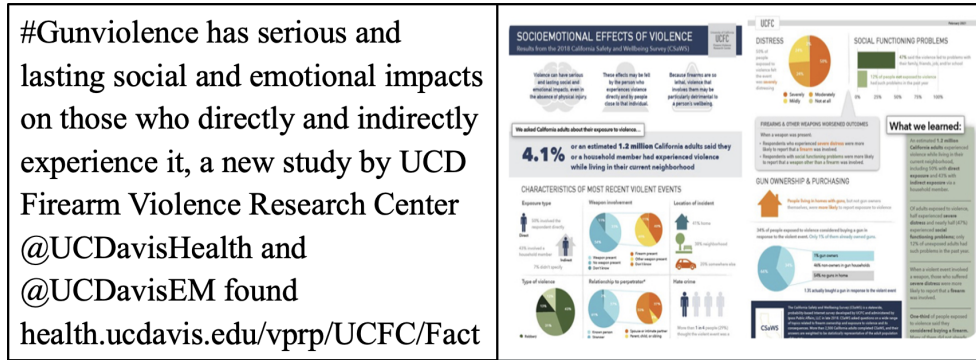
Our work addresses the widely existing issue of political ideology bias in data collection from social media. While this bias problem is not directly related to our main task of modeling modes of persuasion, it is crucial to acknowledge and mitigate bias to ensure the integrity and reliability of the collected data. Further details on this study can be found in our publication [37].

3.4 PersuaCLIP: Image Representation Learning from Tweet Text Supervision

Pre-trained image encoders have become instrumental in developing computational models for various computer vision tasks, particularly when labeled data is limited. In the case of predicting modes of persuasion with the ImageArg dataset, which consists of only 228 annotated samples, the scarcity of labeled data presents a significant challenge. To overcome this challenge, we propose to leverage self-supervised learning techniques to develop pre-trained models using a large amount of unlabeled data. The primary objective is to enable the self-supervised model to acquire persuasion-related knowledge.

Traditional pre-training objectives for image representation learning involve pretext tasks that focus on image transformations such as colorization, jigsaw puzzles, cropping, noise or blur. While these pre-trained models excel in tasks like object detection or segmentation by understanding the *visual content in a scene*, they are not specifically designed to capture the underlying *intent or persuasive elements within the scene*. For the task of predicting persuasion modes, it is essential to go beyond mere object recognition. For instance, consider the example of a statistical chart shown in Figure 13, which is a salient type of visual content related to persuasion. Conventional image encoders, like ResNet pre-trained on ImageNet [43], may not effectively capture such content.

However, we recognize that the accompanying text in tweets contains valuable clues that can aid in predicting the persuasion modes. For instance, in Figure 13, phrases like “a new study” may indicate the utilization of logical reasoning (mode of logos), while references to the “UCD Firearm Violence Research Center” suggest an appeal to credibility (mode of ethos). Based on this observation, we propose using the accompanying text as the self-supervised signal for pre-training an image encoder capable of understanding visual persuasion. By employing the text as a supervision signal during the self-supervised learning process, we aim to train the model to acquire the necessary knowledge to predict the modes of persuasion in tweet images. The objective is for the self-supervised model to learn the relationship between visual and textual cues and develop a deeper understanding of persuasive visual content. This approach enables the model to capture the nuances and subtle



(a) A posted tweet text

(b) An associated posted tweet image

Figure 13: Example of a tweet containing statistical charts: (a) the tweet text uses gun violence to argue for *gun control*; (b) the image makes the argument more persuasive by providing supplementary statistics relating violence to gun ownership in California.

elements associated with persuasion, ultimately enhancing its ability to analyze and predict the modes of persuasion utilized in persuasive images.

In developing a self-supervised model for predicting modes of persuasion, two key design components play a crucial role: model architecture and pre-training objectives. These choices significantly impact the model’s ability to learn and capture the complex relationships between visual and textual cues. In this section, we delve into our decisions regarding these design components for our proposed model *PersuaCLIP*.

3.4.1 Architecture of PersuaCLIP

In our pursuit of an effective model architecture for capturing the interplay between visual and textual cues in persuasive images, we have selected CLIP (Contrastive Language-Image Pre-training) [100]. As discussed in Section 2.3.3, CLIP’s architecture is specifically designed to bridge the semantic gap between images and their associated text. It leverages a transformer-based model that encodes both images and text, enabling a unified understanding of the multi-modal input. Specifically, a GPT-style transformer [103] for the text encoder and Vision Transformer (ViT) [30] for the image encoder.

Furthermore, CLIP benefits from its large-scale pre-training on diverse and extensive datasets, encompassing a wide range of visual concepts and linguistic patterns. This pre-training process enables the model to learn the nuanced connections between visual and textual elements, providing a strong foundation for modeling the persuasive aspects of visual rhetorics. By adopting CLIP as our model architecture, we leverage its state-of-the-art performance, robust multi-modal representation learning, and its ability to capture the intricate relationships between visual and textual elements. With CLIP as our foundation, we aim to develop an effective computational model for understanding and predicting the modes of persuasion utilized in persuasive images.

3.4.2 Pre-training Objectives of PersuaCLIP

To align our pre-training objectives with CLIP, we adopt the same approach of predicting whether an image is paired with a text. The model learns a multi-modal embedding space by jointly training an image encoder and text encoder to maximize the cosine similarity of the image and text embeddings of the real pairs while minimizing the cosine similarity of the embeddings of the incorrect pairings. By using CLIP’s pre-trained weights, we initialize the image and text encoders, benefiting from the learned knowledge and representations. For the pre-training corpus, we collect raw tweets containing both an image and text. To increase the complexity of the matching task and avoid simplistic correlations, we focus on tweets related to the same topic as ImageArg (*i.e.*, gun control). By doing so, the model cannot rely solely on topic similarity to predict the matching task, pushing it to learn more intricate relationships between the paired image and text, specifically related to persuasion.⁹

To introduce a higher level of difficulty in the matching task, we propose masking certain tokens in the tweet text. This masking helps prevent the model from simply selecting the correct label without truly understanding the key relationship between the image and text pair. Our aim is to ensure that the model grasps the underlying connection that goes beyond surface-level cues. For this purpose, we identify and mask what we refer to as “shortcut”

⁹In addition, we also implement a filtering process to remove certain elements from the tweet text that could serve as potential shortcuts for the model to solve the pre-training tasks without truly learning the target knowledge. This filtering process involves removing hashtags, usernames, URLs, and other similar elements.

tokens in the tweet text. These shortcut tokens should be unrelated to persuasion but still informative for predicting whether the image and text are a match. To achieve this, we leverage expert-selected keywords that were used to retrieve topic-relevant tweets. These keywords define specific subtopics within the broader social topic, ensuring that they are related to the content of the tweets rather than their persuasive nature. For instance, in the context of the “gun control” topic, some example keywords include “gun free”, “second amendment”, “progun”, “PrayForOrlando”, and others. By masking these shortcut tokens, we prevent the model from “cheating” by relying solely on the presence of a keyword in the text to match it with the image. Instead, the model is encouraged to focus on the deeper relationship between the image and text that is indicative of persuasion modes. To maintain consistency between training and testing, we have made the decision not to mask any regions in the image during the pre-training phase. Our ultimate goal is to utilize the pre-trained image encoder for predicting the modes of persuasion employed in images. In order to achieve this goal, it is crucial that the image encoder processes the complete image during the training process.

However, the removal of shortcut tokens alone does not guarantee that the model will focus solely on persuasion, as other relationships between the image and text may confound the task. To address this, we decide to use only persuasive tweets for pre-training, further emphasizing the importance of capturing persuasion-related knowledge.

By incorporating these modifications into our pre-training process, we aim to create a more challenging and tailored learning environment for the model. This approach allows us to adapt the learned knowledge from CLIP specifically to the social media context, enhancing the model’s ability to understand and predict the modes of persuasion utilized in persuasive images.

3.5 Experiments

We conduct a series of experiments to evaluate the performance of the proposed Persua-CLIP model in predicting persuasion modes of tweet images. The experiments consist of

two main steps: pre-training the PersuaCLIP model using a tweet corpus and evaluating its performance on the ImageArg dataset.

3.5.1 Pre-training PersuaCLIP

To begin with, we pre-train the PersuaCLIP model using a variation of pre-training objectives. To gather a suitable pre-training corpus, we utilize the TwitterAPI¹⁰ to collect raw tweets that fulfill the following criteria: each tweet contains both an image and text, and the content is related to the topic of gun control. The collection period spans a 10-year timeframe, from 03/30/2013 to 03/08/2023. In total, we obtain 66,126 multi-modal tweets meeting these requirements. To prepare the data for pre-training, we split the dataset into a training set of 60,000 tweets and a validation set of 6,126 tweets. This division allows us to effectively train and evaluate the matching performance of the PersuaCLIP model.

In order to mask the shortcut tokens during pre-training, we use the same expert-selected keywords as mentioned in Section 3.2. Additionally, we construct a set of persuasive tweets by employing the ArgumentText Classify API¹¹. This API allows us to compute the argumentative score of the tweet text, and we retain only those tweets with a score above 0.5. A score above 0.5 indicates a higher likelihood of containing substantial argumentative content, allowing us to prioritize more convincing and influential tweets in pre-training data. However, it is important to acknowledge that the classifier model may not be entirely accurate in all cases and can make errors in their predictions. Since the purpose of pre-training is to expose the model to a wide variety of examples to learn general language representations, the specific threshold used for data selection is less critical. As a result, the initial training set of 60,000 tweets is reduced to 9,240 after applying this filtering criterion. The validation set consists of 930 tweets.

For the pre-training process, we utilize an existing toolkit¹² with specific configurations. The model is trained with a batch size of 64, a learning rate of 1e-6, and a total of 3 epochs. Throughout the pre-training phase, we continuously evaluate the performance of the model

¹⁰<https://developer.twitter.com/en/docs/twitter-api>

¹¹<https://api.argumentsearch.com>

¹²https://github.com/mlfoundations/open_clip

Table 4: Image-text matching performance on tweet corpus.

Model	Image to Text			Text to Image		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	23.93	40.09	47.27	19.26	33.01	39.47
PersuaCLIP	30.49	49.84	58.19	28.44	48.25	56.69

on the validation set. The results obtained demonstrate the superiority of our pre-trained PersuaCLIP model in matching tweet images with their corresponding text compared to the CLIP model, as shown in Table 4. These results serve as evidence that pre-training on a tweet corpus significantly enhances the model’s understanding of social media content.

3.5.2 Evaluation of PersuaCLIP with ImageArg

To ensure the reliability of our results, we acknowledge the limited size of annotated samples in the ImageArg dataset, which consists of only 228 samples. In order to mitigate the impact of this small dataset, we employ a 5-fold cross-validation approach. This methodology helps to reduce the potential bias and variability by randomly dividing the dataset into five subsets. We use 60% of the data for training, 20% for validation, and the remaining 20% for testing in each fold. By averaging the performance metrics across the five folds, we obtain a more representative estimate of the model’s accuracy, precision, recall, F1 score, and AUC.

For preprocessing the tweet texts, we apply the same methods as described in our published work [74], which involve removing Emoji, URLs, Mentions, and Hashtags. The images are resized to a dimension of 224x224. As a baseline model, we utilize ResNet50, pre-trained on ImageNet [43]. In contrast to our published work, where separate binary classifiers were trained for each persuasion mode, in this dissertation, we employ a single classifier layer with three prediction heads sharing weights, enabling multi-label prediction. This approach allows the model to learn the interconnected relationship between the three modes of persuasion. Additionally, no visual augmentation techniques are used during training, and a smaller learning rate is applied with an extended number of training epochs. As a result, the experimental results of the ResNet baseline in this dissertation may differ from previous

Table 5: Prediction performance on modes of persuasion with ImageArg (with its standard error). The input of models is the tweet image.

Task	Model	Precision	Recall	F1	AUC
Logos	ResNet50	69.91 \pm 16.58	54.11 \pm 18.87	58.85 \pm 14.05	80.57 \pm 6.08
	CLIP	77.24 \pm 8.08	66.58 \pm 9.35	70.80 \pm 5.28	89.29 \pm 4.64
	PersuaCLIP	75.82 \pm 8.03	62.02 \pm 7.40	67.97 \pm 6.64	88.77 \pm 4.08
	PersuaCLIP w/ masked tweets	76.88 \pm 9.28	66.19 \pm 8.51	70.65 \pm 7.13	88.44 \pm 3.60
	PersuaCLIP w/ persuasive tweets	74.17 \pm 8.97	59.72 \pm 7.08	65.94 \pm 6.98	88.66 \pm 4.26
Pathos	ResNet50	73.32 \pm 6.16	68.50 \pm 10.28	70.32 \pm 7.14	80.66 \pm 9.31
	CLIP	72.21 \pm 11.11	75.60 \pm 9.12	72.60 \pm 4.89	82.64 \pm 2.21
	PersuaCLIP	71.83 \pm 9.82	73.86 \pm 6.62	71.94 \pm 3.18	82.81 \pm 2.30
	PersuaCLIP w/ masked tweets	72.57 \pm 10.43	75.71 \pm 9.78	72.94 \pm 5.28	82.59 \pm 2.91
	PersuaCLIP w/ persuasive tweets	73.69 \pm 9.86	78.36 \pm 7.24	75.05 \pm 4.01	82.49 \pm 2.48
Ethos	ResNet50	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	69.93 \pm 9.83
	CLIP	80.00 \pm 40.00	15.78 \pm 8.09	26.32 \pm 13.39	83.47 \pm 9.01
	PersuaCLIP	80.00 \pm 40.00	17.56 \pm 13.12	27.81 \pm 18.58	86.74 \pm 6.28
	PersuaCLIP w/ masked tweets	80.00 \pm 40.00	17.56 \pm 13.12	27.81 \pm 18.58	87.11 \pm 5.86
	PersuaCLIP w/ persuasive tweets	40.00 \pm 48.99	10.22 \pm 15.50	15.43 \pm 22.25	84.58 \pm 6.40

reports.

Due to the limited size of the training data, we train a shallow classifier head on top of the pre-trained image or text encoder of CLIP or PersuaCLIP while keeping the encoders frozen. We conduct experiments using the image modality or the image-text multi-modality as input. For the multi-modality input, we concatenate or average the image and text embeddings before feeding them into the classifier layers. The networks are optimized using the Adam optimizer with a learning rate of 1e-4 and a batch size of 16.

The experimental results with image input are presented in Table 5. We observe that both CLIP and PersuaCLIP models outperform the ResNet baseline, affirming the effectiveness of utilizing accompanying text as a self-supervised signal (hypothesis **H1**). The improvements in predicting Logos or Ethos, as reflected in precision, recall, F1, and AUC, are statistically significant, exceeding the standard error. On the other hand, for Pathos, CLIP and PersuaCLIP models exhibit better recall while maintaining nearly the same precision. This suggests that the incorporation of accompanying text as a self-supervised signal has a more substantial impact on recall, enabling the models to better identify instances

Table 6: Prediction performance on modes of persuasion with ImageArg (with its standard error). The input of models is both the image and the text. Comparing with models using image input, the improvement of F1 or AUC is marked in green and the drop is marked in red.

Task	Model	Precision	Recall	F1	AUC
Logos	CLIP	71.68 \pm 10.62	61.80 \pm 13.38	65.60 \pm 9.76	89.01 \pm 5.52
	PersuaCLIP	78.51 \pm 7.77	68.76 \pm 9.33	73.04 \pm 7.56	88.81 \pm 5.88
	PersuaCLIP w/ masked tweets	77.88 \pm 8.84	69.81 \pm 7.59	73.42 \pm 7.03	89.07 \pm 5.61
	PersuaCLIP w/ persuasive tweets	80.69 \pm 8.24	66.40 \pm 10.03	72.48 \pm 8.01	88.41 \pm 6.34
Pathos	CLIP	73.93 \pm 7.96	75.92 \pm 8.91	74.23 \pm 4.77	83.14 \pm 3.73
	PersuaCLIP	74.17 \pm 7.69	75.34 \pm 8.76	74.29 \pm 5.81	84.17 \pm 3.69
	PersuaCLIP w/ masked tweets	74.40 \pm 6.98	72.97 \pm 10.10	73.07 \pm 5.64	83.90 \pm 3.92
	PersuaCLIP w/ persuasive tweets	72.96 \pm 7.93	72.97 \pm 10.10	72.31 \pm 5.97	84.47 \pm 4.69
Ethos	CLIP	30.00 \pm 40.00	6.22 \pm 8.12	9.71 \pm 12.20	83.68 \pm 7.00
	PersuaCLIP	20.00 \pm 40.00	2.22 \pm 4.44	4.00 \pm 8.00	87.55 \pm 4.76
	PersuaCLIP w/ masked tweets	40.00 \pm 48.99	6.22 \pm 8.12	10.67 \pm 13.73	87.50 \pm 3.81
	PersuaCLIP w/ persuasive tweets	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	85.30 \pm 5.53

that belong to each persuasion mode.

In most cases, PersuaCLIP and its variations perform better than CLIP, with the exception of predicting Logos. Interestingly, masking tokens in tweets or incorporating persuasive tweets improves the performance of PersuaCLIP in certain scenarios. For instance, PersuaCLIP with persuasive tweets shows promise in effectively capturing the Pathos mode. However, the differences between the variations rarely significant. Notably, the Ethos mode proves to be more challenging for all models, with limited performance across the board. One contributing factor is the scarcity of annotated examples for Ethos compared to Logos or Pathos, making the learning process more demanding.

We further conduct experiments using both image and text as input, and the results are presented in Table 6. Comparing with Table 5, we observe that the multi-modal performance generally outperforms the single-modality models. For example, PersuaCLIP’s F1 score increases from 67.97 to 73.04 for logos prediction when incorporating both image and text. This improvement can be attributed to the accompanying text’s ability to express the logic and reasoning conveyed in the logos images. Comparing different models within Table 6, we

find that PersuaCLIP consistently outperforms CLIP for both Logos and Pathos modes of persuasion. However, Ethos remains a challenging mode for all models.

These findings highlight the effectiveness of PersuaCLIP and its variations in predicting modes of persuasion using the ImageArg dataset, with multi-modal approaches yielding improved performance compared to single-modality models. Nevertheless, Ethos prediction remains a difficult task due to the limited availability of annotated examples.

3.5.3 Limitations

While our study provides valuable insights and advancements in predicting persuasion modes of tweet images using the PersuaCLIP model, there are certain limitations that should be acknowledged:

1) Limited annotated dataset: The size of the annotated dataset in ImageArg is relatively small, consisting of only 228 samples. This limited dataset size may affect the generalizability of the model’s performance and introduce a certain degree of uncertainty in the reported results. It is important to acknowledge that the conclusions drawn from this dataset may not fully capture the variability and complexity of persuasion modes in tweet images.

2) Imbalanced class distribution: The distribution of persuasion modes within the annotated dataset may be imbalanced, with certain modes having fewer instances compared to others. This could potentially impact the model’s ability to accurately predict less-represented persuasion modes. Care should be taken when interpreting the performance of the model for specific modes with limited instances.

3) Generalizability to other topics: Our study specifically focuses on the topic of gun control, and the PersuaCLIP model is trained and evaluated on tweets related to this topic. The generalizability of the model’s performance to other social topics or domains may vary, as the characteristics and patterns of persuasion modes could differ across different contexts. Future research should explore the applicability of PersuaCLIP to a wider range of topics to assess its robustness and adaptability.

These limitations provide valuable insights into the potential constraints and challenges of our study. Future research efforts should address these limitations by considering larger

and more diverse datasets and extending the evaluation to different social topics to enhance the generalizability and applicability of PersuaCLIP in real-world scenarios.

3.6 Chapter Summary

In this chapter, our focus is on analyzing persuasive elements in multi-modal media, with a specific emphasis on modes of persuasion in images. To facilitate this analysis, we have created a new annotated dataset (*i.e.* ImageArg), and we have developed computational models (*i.e.* PersuaCLIP) to predict the modes of persuasion present in the images. We encountered several challenges during the collection and annotation of the dataset, including the need to adjust existing annotation schemes and devise a novel taxonomy. Through our efforts, we have discovered that the modes of persuasion can be effectively adapted for the analysis of persuasive images, thus supporting our initial hypothesis (**H1**). By leveraging self-supervised training on a large-scale collection of multi-modal tweets and utilizing the accompanying text as a form of supervision signal, we have enhanced our model’s understanding of the persuasion modes depicted within the images. These findings strongly support our hypothesis **H1** that the textual content accompanying the images serves as a valuable self-supervised signal for classifying the persuasion modes. Despite the challenges in representing images, our research contributes to the understanding of persuasive elements in multi-modal media and provides insights into the rhetorical strategies employed.

4.0 Detecting Atypicality

4.1 Introduction

In this chapter, we investigate detecting atypicality in persuasive images, called *persuasive atypicality*. Visually creative images, such as advertisements or public service announcements, may purposefully contain atypical portrayals of objects as a rhetorical way for attracting viewers’ attention [47]. In the marketing and communications research community, atypicality has gained attention because of its importance to understanding the persuasiveness and rhetoric of visual media [82, 72, 143]. However, detecting persuasive atypicality is under-explored. Prior work focuses on detecting atypical objects in real-world images [136], such as diverse defects [10] and out-of-context objects and scenes [23, 108]. Most prior studies investigate atypicality that is (1) physically created in the real world, rather than generated with computer graphics; and (2) predominantly accidental and certainly not aiming to convey meaning or persuade an audience to take a certain action.

Detecting persuasive atypicality is more challenging for intelligent systems. First, atypicality may involve metaphorical object transformations or intentionally surprising composed objects (*e.g.* a kiwi inside an apple). Second, the atypicality transformation types are diverse and not limited to a specific set of categories, as they are in Wang *et al.*’s work [136], in which atypical objects are all from PASCAL VOC [32]. Third, unpacking them may require common-sense reasoning. For example, Figure 14a is an atypical advertisement for a beverage. It is unusual for a pig to wear a bridal veil even though the pig and veil are both normal objects. The ability to detect this type of purposefully atypical objects and understand their roles in conveying the intent of the image is necessary for an intelligent system to reason about information in persuasive media. In this work, we propose to model implicit knowledge of contextual compatibility by self-supervised learning in order to detect persuasive atypicality.

Our hypothesis **H2** is that persuasive atypicality can be detected by checking the compatibility between each possibly atypical object and the rest of the image as context. For

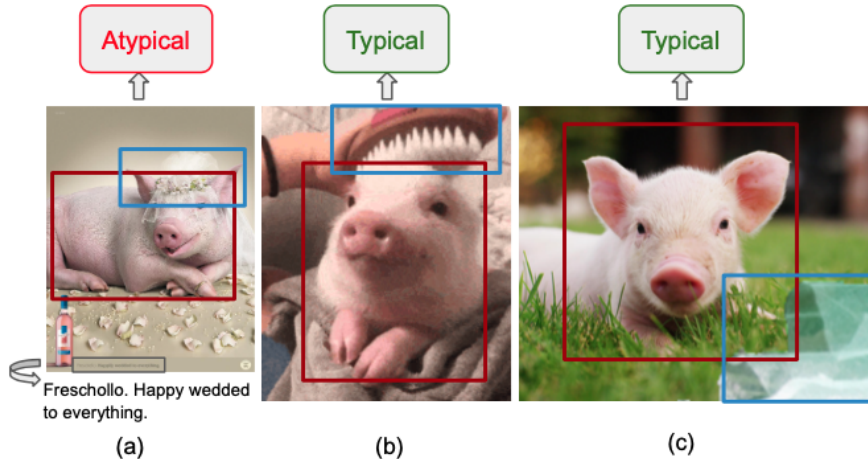


Figure 14: These images illustrate the importance of object interactions and their spatial relative position for atypicality detection. (a) Pig wearing a bridal veil is atypical; (b) If a handled brush instead of a veil is on top of the pig’s head, then the image is typical; (c) If the veil’s location is different, the image may also be typical.

example, in Figure 14a, the pig is not compatible with its context (a bridal veil on its head), and the veil is also not compatible with its context — on a pig’s head. We propose an self-supervised approach by using reconstruction losses of masked regions. We expect that a self-supervised model trained on masked region reconstruction could learn enough implicit knowledge of contextual compatibility; this pre-trained model may then be used to detect atypical images.

Additionally, we hypothesize that the interactions between objects and their spatial relative positions play a key role in detecting atypicality (**H2**). If it were a handled brush instead of a bridal veil over the pig’s head (Figure 14b), or if the veil were placed at another location instead of on top of the pig’s head (Figure 14c), the image would no longer be atypical. In order to better interpret object-object spatial interaction, we propose a new method to compute the attention weights between key-query regions of our transformer-based models.

Finally, we explore the possibility to only use the textual modality as input, *i.e.* object classes in natural language, instead of the visual images when modeling the contextual com-

patibility. In Figure 14a, knowing that there is a “pig” and a “bridal veil” and their spatial relationship may be helpful to conclude that the image is atypical, instead of knowing exactly what that pig or veil look like. The potential advantage of the vision-to-text translation is to enforce the model focusing on the semantic meaning rather than the concrete visual content.

4.2 Our Approaches

We define atypicality detection as a binary classification task: for a given image, our model aims to predict whether the image is atypical or not. We first present our unsupervised atypicality detection system, which leverages masked region reconstruction as the pretext task, and learns implicit knowledge of contextual compatibility from large-scale unlabeled data. The reconstruction losses of masked regions are the clue for predicting atypicality of a test image. We then introduce our Relative-Spatial Transformer which extends the self-attention layer to explicitly model relative position information separately from visual features.

4.2.1 Masked Region Reconstruction

Figure 15a shows an overview of our approach. An image I is represented by a set of regions $R = \{(v_1, p_1), (v_2, p_2), \dots, (v_n, p_n)\}$, where v_i could be region i 's visual feature vector, pixel matrix, class labels, etc., and p_i is the positional information. Our hypothesis is that if an image is atypical, the objects appearing in it would not be compatible with each other, thus it would be hard to reconstruct a masked region from image context. We first pre-train a model to reconstruct a region from context using normal cases, then use it to detect atypicality in new test images.

For the pre-training process, we take inspiration from masked language modeling (*e.g.* BERT [28]) and cross-modality representation learning (*e.g.* LXMERT [127]). The model is trained to reconstruct the masked regions given the remaining regions, on many general, normal images (which could potentially contain a small proportion of atypical cases). Different

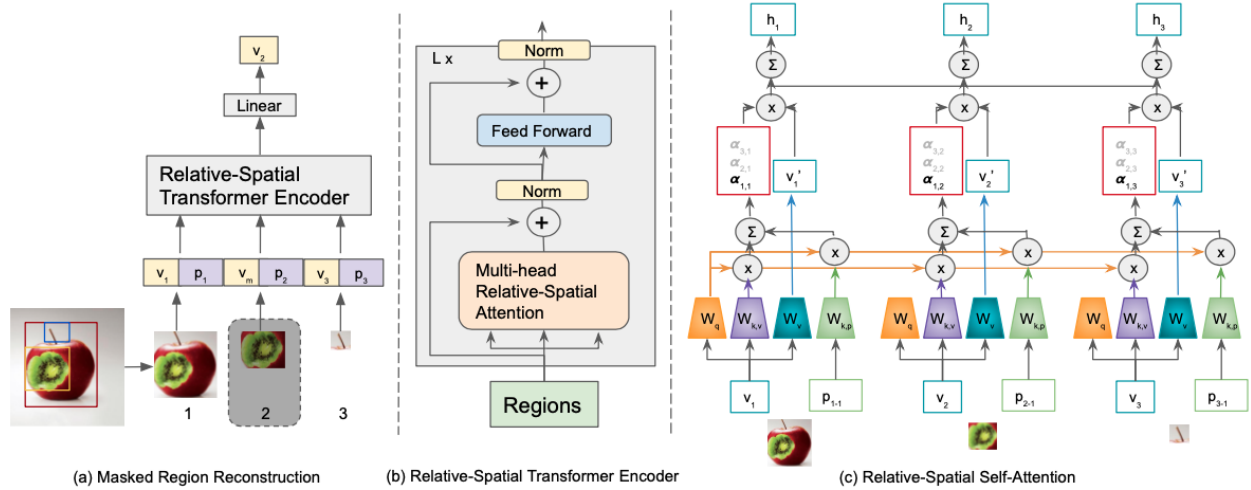


Figure 15: Model overview. (a) A set of regions extracted from the image are the input to the Relative-Spatial Transformer Encoder. The model is trained for reconstructing the visual feature of the masked region 2. (b) The architecture of Relative-Spatial Transformer Encoder. The key difference from the standard Transformer Encoder is the attention computation. (c) The mechanism for computing Relative-Spatial Self-Attention. This scheme shows the case when region 1 is the query.

from BERT or LXMERT, which aims to learn a language or visual-language representation, our model aims to learn the common co-occurrences and typical spatial relationship between objects.

At test time, we mask each region in the image and compute the reconstruction loss. We compute the average loss of all regions as a clue for predicting atypicality. We use average rather than maximum loss because if an image is atypical, the masked region reconstruction loss is high not only when an atypical object is masked, but also when its surrounding object is masked since it is also hard to reconstruct a normal object from an atypical context.

4.2.2 Relative-Spatial Transformer

Our model extends the transformer architecture [128]. Since the transformer is permutation-invariant, a positional encoding is necessary to provide the order information of the sequential input. For work which represents the image by a set of regions of interest, a common way is to embed the bounding-box coordinates of each region and potentially the fraction of image area covered [22, 127, 75]. Then the summation of the visual embedding and the positional embedding of the region is used as the input features for the transformer [15, 127, 30]. However, this technique has two weaknesses. First, when computing attention weights with these input vectors, the visual feature and positional information share the same projection weight without any distinction, therefore the model cannot flexibly adjust the importance of region visual and position. Explicitly modeling relative position information separately from other inputs (e.g. features) extends the self-attention mechanism to efficiently consider spatial relationship between each query-key pair [114, 104, 9, 149]. Second, the positional embedding represents the absolute coordinate of the region, however, it is the *relative* spatial relationship between the masked and the context region which matters for detecting atypicality (e.g. is the veil above or below the pig?). Experimentation in machine translation [114] and music generation [48] suggested that using relative positional embeddings results in significantly better accuracy. In order to overcome both weaknesses, we propose the Relative-Spatial Transformer which (1) computes the visual-visual interaction and visual-position interaction separately, and (2) is shift-invariant, similar to convolutions but unlike a standard transformer.

Our approach follows previous ideas that define 2D relative position embeddings by the relative distance between the position of the query and key pixel [104, 9], except that our relative position embedding is at the region level. Besides, we can add overlapping area information to the relative spatial feature between two regions, which a pixel-level representation cannot. Kant *et al.* also consider relative spatial relationship between object regions, but they transform spatial relationship into twelve categories and then apply the adjacency matrices as an additional attention mask on their base model architecture [55]. Therefore, they only consider the relative spatial direction and ignore the concrete relative

distance between pairwise objects, which loses essential information compared to our method. Another weakness is their spatial relationship categories do not have full coverage, *e.g.* the spatial relationship between two non-overlapped objects far from each other is ignored.

Our proposed Relative-Spatial Transformer (RST) follows the same architecture as the transformer (T) [128] except for a new way for computing the multi-head self-attention layer, as shown in Figure 15. The attention weight of the query region i and key region j is computed as:

$$A_{i,j}^{rel} = V_i^T W_q^T W_{k,V} V_j + V_i^T W_q^T W_{k,P} P_{j-i} \quad (1)$$

where V_i and V_j are visual features of regions i and j ; W_q is the projection weight of the query region visual feature; $W_{k,V}$ and $W_{k,P}$ are respectively the key region’s projection weights of visual features and relative positions; and P_{j-i} is the relative position of region j with respect to region i . The first term computes the interaction between the query and key visual content; the second term computes the interaction between the query visual content and the relative position of the key region. The summation of both terms shows the importance of the key region to the query region. Then we compute the normalized attention weight $\alpha_{i,j}^{rel}$ as a softmax layer over $A_{i,j}^{rel}$ for all possible key regions. The last hidden layer of region i is computed as:

$$h_i = \sum_j \alpha_{i,j}^{rel} W_v V_j \quad (2)$$

where W_v projects the value region’s visual feature.

The reconstruction loss of the masked region i is computed as the mean squared error (*i.e.* squared L2 norm) between the input visual feature v_i and the last hidden layer h_i of the encoder:

$$L_i = \|v_i - h_i\|_2^2 \quad (3)$$

For computing the relative position of j with respect to i , we compute the x-axis and y-axis distance of the top-left and bottom-right corners of the two bounding boxes:

$$P_{j-i} = [x_j^l - x_i^l, y_j^t - y_i^t, x_j^r - x_i^r, y_j^b - y_i^b] \quad (4)$$

where (x_i^l, y_i^t) is the coordinate of the left-top corner of region i , (x_i^r, y_i^b) is the coordinate of the right-bottom corner of region i ; similarly with region j . We also explore adding Intersection-over-Union area between region i and j as an additional relative positional feature.

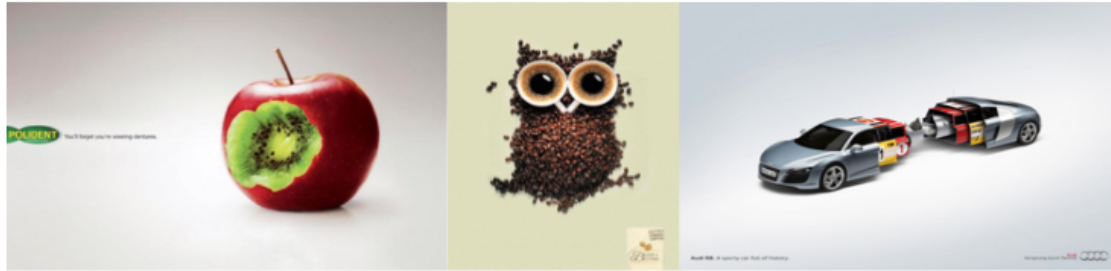
4.3 Experiments

In the subsequent experiments, we first evaluate our contextual compatibility modeling approach on detecting persuasive atypicality in the Ads dataset [143]. Our experiments show that Relative-Spatial Attention leads to an improvement across a diverse array of atypicality sub-categories. Then, to understand the labelling requirement of the task, we compare our unsupervised contextual compatibility approaches with supervised models trained on the atypical/typical labels. Finally we compare visual versus semantic compatibility for examining the possibility for representing the image context by text.

4.3.1 Setup

Data. We evaluate our method on the Ads dataset with their annotations on atypicality [143]. Since each image is annotated by one or multiple annotators, we set a rule for deciding the atypical/typical label if annotators do not agree with each other. In particular, we consider an ad atypical if any annotator labels it as atypical. We use the *ifany* rule because some atypical cases are subtle, subjective or need background knowledge, thus any annotator providing the atypical label is cause to believe the image is not quite typical. Under this labeling rule, there are 2,285 atypical ads and 1,643 typical ads. For the self-supervised training, we use all ads except for those 3,928 with atypicality labels, resulting a set of around 60k images. For supervised training and for testing, we randomly split these atypical/typical images using a 7:1:2 ratio for train:val:test sets.

In addition to the binary label saying whether the image is atypical or not, one or several atypicality categories are annotated for each atypical image. The eight categories of



(a) Apple with kiwi texture

(b) Owl made in coffee

(c) Auto racing in car



(d) Woman without mouth

(e) Deer with hand horn

(f) Bent arm



(g) Beer deformed as a hockey player

(h) Cigarettes as bullets

Figure 16: Atypical object transformations in the Ads dataset [143].

atypicality are defined based on object transformations:

- 1) Texture Replacement 1 (**TR1**): Objects' texture borrowed from another object, *e.g.* kiwi inside apple, Figure 16a.
- 2) Texture Replacement 2 (**TR2**): Texture created by combining several small objects, *e.g.* owl from beans, Figure 16b.
- 3) Object Inside Object (**OIO**), *e.g.* auto racing in car, Figure 16c.
- 4) Object with Missing Part (**OMP**), *e.g.* woman without mouth, Figure 16d.
- 5) Combination of Parts (**CP**): Object composed by parts from different objects, *e.g.* deer

Table 7: The annotated description of atypical objects for each category [143].

Category	Annotation Template
TR1	The object which has a new texture is [OBJ1] and the new texture is [OBJ2].
TR2	The object which has a new texture is [OBJ1] and the objects that have created the texture are [OBJ2].
OIO	Objects which are inside are [OBJ1] and the objects that are outside are [OBJ2].
OMP	The objects which have missing parts are [OBJ1].
CP	The objects that have created new object are [OBJ1].
SDO	The objects that have been transformed are [OBJ1] and the transformations are [OBJ2].
LDO	The liquid which has been deformed is [OBJ1].
OR	The object which is placed in the context of another object is [OBJ1] and the object which is replaced by another object (expected object) [OBJ2].

head with hand horn, Figure 16e.

6) Solid Deformed Object (**SDO**), *e.g.* human arm bent, Figure 16f.

7) Liquid Deformed Object (**LDO**), *e.g.* beer as player, Figure 16g.

8) Object Replacement (**OR**): The whole object appearing in the context normally associated with another, *e.g.* cigarettes placed in the context where bullets occur, Figure 16h.

In addition, if an image contains an atypical object but it cannot be assigned to any of the aforementioned categories, then it is annotated as “Others”. Regarding the statistics on atypical categories, Object Replacement is the most prevalent category, followed by Combination of Parts, while Object with Missing Part is the least frequent. The raw inter-rater agreement for these categories ranges from 0.41 (for Texture Replacement 1) to 0.58 (for Liquid Deformed Object), indicating reasonable agreement among the annotators [143].

Moreover, atypical images are also annotated with a textual description about the concrete atypical objects by a fixed template. For example, for the category of TR1, a description is collected in the form of “The object which has a new texture is [OBJ1] and the new texture is [OBJ2]”. The templates for each category is shown in Table 7. These atypicality categories and description of objects in atypical images enable us for conducting some fine-grained analysis and providing more insights.

Input Representations. We use Faster R-CNN [105] pre-trained on Visual Genome

[58] for extracting the visual features [6]. Faster R-CNN itself uses ResNet-101 [43] pre-trained for classification on ImageNet [107]. We take the features of each detected object as the visual representation of the corresponding region. We select a fixed number of objects (36) by sorting detections by confidence score. Each region is represented by its bounding-box coordinates and its 2048-dimensional region-of-interest (RoI) features.

Self-supervised Training and Testing. Following BERT [28], we mask 15% of regions in each sequence at random during training. All masked regions are replaced by a trainable vector with the same dimension as the RoI feature. The spatial information of the masked region is given. We use a batch size of 128 and train for 20 epochs with learning rate of 1e-3. For testing, we mask one region with the learned vector at a time, then compute the average reconstruction loss of all regions. The higher the loss, the more likely the image is atypical. We compute the ROC-AUC score as the evaluation metric since it measures model performance across all possible classification thresholds, by reporting the probability the model ranks a random atypical example higher than a random typical one.

Model Size. We denote the number of layers (*i.e.*, transformer blocks) as L , the hidden size as H , and the number of self-attention heads as A . We primarily report results on the model with $L=1$, $H=768$, $A=8$ ¹.

Baseline Models. We consider two baselines, Auto-encoder and One-Class SVM, since they are standard methods for detecting abnormality and outliers [5, 67]. For the **Auto-encoder**, we implement the same encoder as DCGAN’s discriminator and DCGAN’s generator as the decoder [101], using the hyperparameters in [101]. The loss is L2 error between input and generated images. However, we make an interesting observation that atypicality relates to image complexity in a potentially counter-intuitive way. We found strong correlation between atypical images and relatively plain backgrounds, likely because ad designers of atypical images want to make sure the image is plain enough for the audience to notice the atypicality. Images with uniform background are more easily reconstructed while images with plenty of objects are harder. Further, images with more pixels tend to contain more information to be compressed and reconstructed. To ensure the auto-encoder captures atypi-

¹There are no extra trainable parameters in RST compared to T. In Figure 15c, $W_{k,v}$ is extra parameters of size of d_p*d_v (dimensions of position p & visual v vectors), but unlike RST, T requires trainable parameters of size $d_p * d_v$ for projecting p to the same dimension as v for the summation.

cality rather than complexity, we need to normalize for image complexity. We first preprocess all images by resizing them to a fixed number of pixels (64*64). We also measure image complexity (IC) as the average of horizontal and vertical gradient of pixels ($IC = avg(I_x^2 + I_y^2)$ where I_x and I_y are respectively the horizontal and vertical gradient). Then we divide the auto-encoder reconstruction loss by IC . In addition, to force the auto-encoder model to learn an effective encoder and decoder, we limit the dimension of the middle hidden layer to 2048 which is much smaller than the input image dimension (3*64*64). For the **One-Class SVM** model, we represent each image by the average of its 36 RoI feature vectors. Then we fit the One-Class SVM model² with default settings on the training images.

Evaluation Metrics. We report the micro-average for evaluation metrics to assess the overall performance of our models across all instances in the dataset, treating each instance equally, regardless of its category or class. The micro-average is particularly useful when there is an imbalance in class distribution, as it gives equal importance to every instance and provides a comprehensive measure of overall model performance.

4.3.2 Unsupervised Persuasive Atypicality Detection

The experimental results of our unsupervised contextual compatibility approaches are shown in Table 8. To gain insights on the impact of different types of persuasive atypicality on the detection result, we also report the model performance on the eight atypicality categories separately. Experimental results show that our approaches significantly outperform baseline models overall (MICRO AVE) and for Combination of Parts (CP), Object Replacement (OR), Others (with p-value < 0.05 by paired Student’s t-test). While Transformer (T) is an existing architecture, and Relative-Spatial Transformer (RST) is our new design, neither has been used for atypicality detection before. These results demonstrate that our approach of checking for contextual compatibility is effective for detecting persuasive atypicality. T outperforms the simpler baselines significantly, but RST achieves the best results overall. By looking into each category, RST leads to an improvement across a diverse array of atypicality types. Notably, RST demonstrates improvements over T that exceed the

²<https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM>

Table 8: Performance of unsupervised models on the Ads dataset for each atypicality category (TR1, TR2, OIO, *etc.*) and the micro average. We report the best AUC with its standard error, and p-value with respect to the best performing model for others by paired Student’s t-test. Our model (RST/T) is starred if significantly better than the best baseline ($p < .05$).

Methods	TR1	TR2	OIO	OMP	CP
Auto-Encoder	54.67 _{p=.03}	63.28 _{p=.41}	38.79 _{p=.00}	52.98 _{p=.09}	57.78 _{p=.02}
One-Class SVM	64.81 _{p=.50}	68.27 _{p=.98}	59.36 _{p=.17}	65.81 ± 5.5	54.21 _{p=.00}
Transformer (ours)	62.66 _{p=.09}	60.72 _{p=.03}	63.07 _{p=.16}	42.52 _{p=.00}	69.18* _{p=.54}
RS Transformer (ours)	67.50 ± 3.8	68.37 ± 4.1	67.31 ± 5.4	55.18 _{p=.03}	71.26* ± 3.7
Methods	SDO	LDO	OR	Others	MICRO AVE
Auto-Encoder	56.62 _{p=.05}	56.05 _{p=.23}	48.57 _{p=.00}	50.99 _{p=.01}	52.52 _{p=.00}
One-Class SVM	65.12 _{p=.40}	54.43 _{p=.07}	56.31 _{p=.04}	54.23 _{p=.04}	58.82 _{p=.02}
Transformer (ours)	63.71 _{p=.22}	61.63 _{p=.53}	64.05* ± 2.9	63.68* ± 3.4	62.86 _{p=.36}
RS Transformer (ours)	68.67 ± 4.8	63.99 ± 4.4	61.84 _{p=.37}	59.68 _{p=.14}	64.32* ± 2.0

standard error (4.84 and 7.65, respectively) for two types of Texture Replacement (TR1, TR2) and by more than 4 points for Object Inside Object (OIO). The atypicality in these categories predominantly arises from unusual spatial relationships between normal objects, involving object compositions. These findings highlight the efficacy of the RST model in capturing and understanding the complex spatial interactions contributing to atypical visual representations, especially in the aforementioned categories. Object with Missing Part (OMP) is the only atypicality category for which the baseline model (One-Class SVM) is better than our approaches. This is because this type of atypicality only comes from a single object without any complex interaction with surrounding objects.

Error analysis. We qualitatively show several cases where the One-Class SVM fails (Figure 17a - d) or both the baseline and our models fail (Figure 17e - h). One-Class SVM fails when atypicality involves composition of normal objects (e.g., cream on top of alcohol bottle), while our transformer models (especially RST) detect this atypicality by learning context via self-supervised training and show large gains. However, our model fails to capture metaphoric similarity: Figure 17e and 17f look typical at first, but shadow versus puma, surfboard versus brand make them atypical. It also fails to interpret symbolic

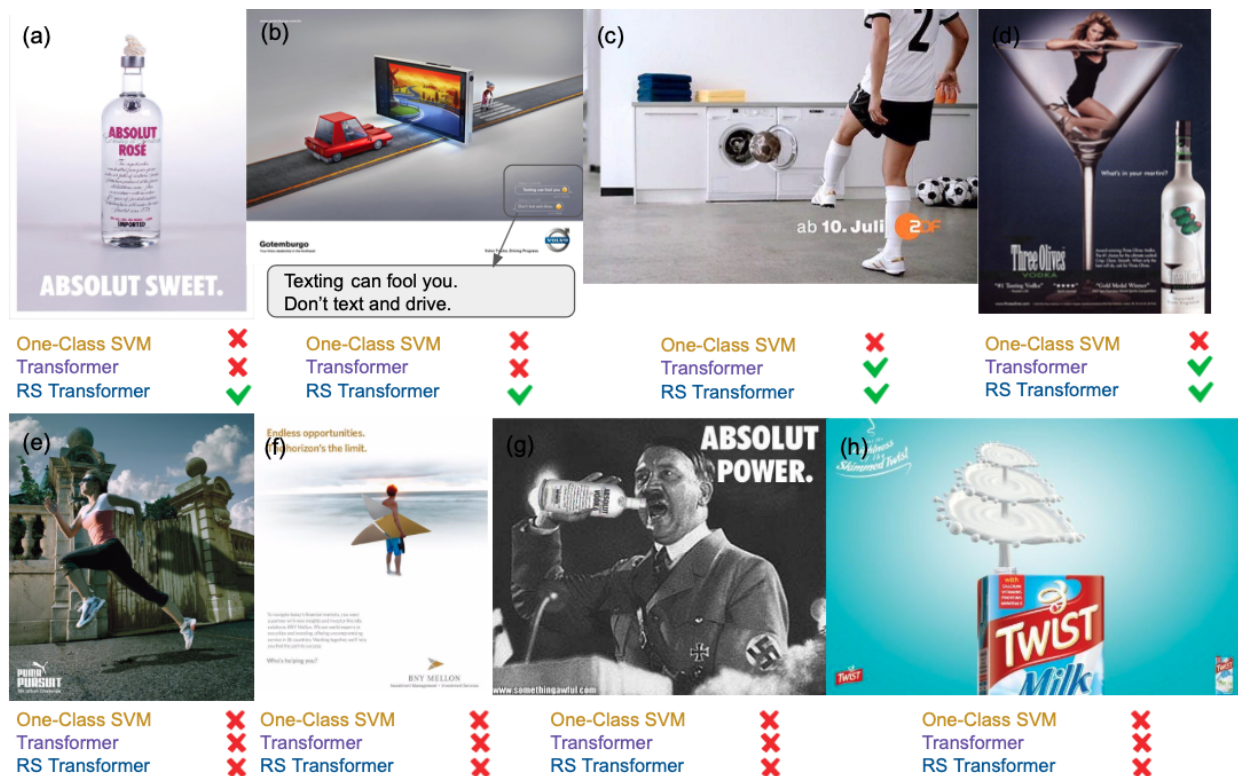


Figure 17: Detection results by the baseline and our models for selected images from the Ads dataset.

Table 9: Ablation study of layer number of encoder and relative positional feature. We include Transformer - L1 and RS Transformer - L1 results from Table 8 for direct comparison with different encoder layer numbers and relation position features. The micro average AUC scores are reported.

Methods	MICRO AVE
Transformer - L1	62.86
Transformer - L4	64.14
RS Transformer - L1	64.32
RS Transformer - L4	64.39
RS Transformer - L1 - IoU	64.99

meanings: vodka is held like a microphone by Hitler who is a symbol of power in Figure 17g. Moreover, in Figure 17h, the milk takes on the appearance of a fountain, which is a completely different and unexpected form for milk to be portrayed in. Models fail to capture this rare and imaginative representation. Thus, typicality judgment requires more fine-grained visual features, and knowledge of historical figures.

Ablation. To see the impact of the number of transformer blocks (model depth), we conduct an ablation study on the layer number (L). Considering the variation of relation position features, we add an additional feature, Intersection-over-Union area (IoU), to the previous relative coordinates. Results are shown in Table 9. We find that the deeper Transformer greatly improves over a shallow Transformer, while Relative-Spatial Transformers are less sensitive to depth. In addition, we observe that a shallow RS Transformer is competitive against a deep Transformer, suggesting that the proposed RS Transformer is more efficient. We also observe that adding the area overlap (IoU) feature slightly improves performance.

4.3.3 Labelling Requirement for the Detection

Models. To understand the labelling requirement for detecting atypicality, we compare our unsupervised contextual compatibility approaches with supervised models trained on the atypical/not labels. We use the same Transformer and RS Transformer architectures for fair comparison. We also add a supervised baseline model which is trained only on the RoI features (each image is represented by the average of all regions-of-interest features).³ For transformers, the output layer is an average pooling over the last hidden layer followed by a simple 2-layer neural network for predicting the atypicality label. For the RoI baseline, the input image features feed directly to the output layer which is the same 2-layer network.

Results. Table 10 shows the comparison of unsupervised and supervised approaches. We find that our unsupervised approaches achieve comparable performance to the supervised approaches, which highlights that even with labeling the task is still difficult. This also demonstrates the effectiveness of our proposed contextual compatibility method. When looking into each atypicality category, we observe the unsupervised RS Transformer outperforms

³The input features are the same as the One-Class SVM baseline. This baseline is conceptually similar to the approach in Ye’s work [143] except that they use VGG16 for extracting the image features.

Table 10: Experimental results on the Ads dataset. We include unsupervised results from Tab. 8 for direct comparison with supervised performance. AUC scores for each atypicality category and the micro average are reported. AE: Auto-encoder; SVM: One-Class SVM; T: our proposed method with Transformer architecture; RST: our proposed method with Relative-Spatial Transformer architecture; RoI: only using RoI features as input.

Methods	TR1	TR2	OIO	OMP	CP	SDO	LDO	OR	Others	AVE	
unsup	AE	54.67	63.28	38.79	52.98	57.78	56.62	56.05	48.57	50.99	52.52
	SVM	64.81	68.27	59.36	65.81	54.21	65.12	54.43	56.31	54.23	58.82
	T	62.66	60.72	63.07	42.52	69.18	63.71	61.63	64.05	63.68	62.86
	RST	67.50	68.37	67.31	55.18	71.26	68.67	63.99	61.84	59.68	64.32
sup	RoI	66.40	65.80	60.13	56.82	63.77	67.41	62.67	62.98	59.41	62.85
	T	66.11	63.16	63.37	64.07	66.55	71.58	70.21	66.03	62.21	65.58
	RST	65.56	64.00	62.20	53.43	70.80	71.11	75.37	67.07	65.59	66.75

the supervised RST on those atypicality transformations which involve more object-object interaction, *e.g.* Texture Replacement 1 or 2, Object Inside Object, Combination of Parts. This is expected because RST efficiently learns contextual compatibility knowledge from the large-scale normal images with the RS self-attention mechanism which is designed for precisely modeling spatial relationship between objects. In addition, the RS Transformer overall outperforms the original Transformer for the supervised setting as well.

4.3.4 Visual versus Semantic Compatibility

We next consider different possibilities for representing the image context, namely checking visual versus semantic compatibility. Our previous experiments use Faster R-CNN RoI features which represent the visual content of the region and then learn compatibility from them. We now consider using the class labels predicted by Faster R-CNN as the semantic features of the region and then we use the same model for learning semantic compatibility.

Training. For unsupervised training with transformer-based models, the input is a sequence of class labels with the bounding-box coordinates of regions ordered by the detection confidence score. Similarly with visual features, we mask one (or several during the training)

Table 11: Comparison of Faster R-CNN RoI visual feature (VF) and predicted class label (CL). AUC for each atypicality category and micro ave are reported, with best AUC per column bolded. T: our proposed method with Transformer architecture; RST: our proposed method with Relative-Spatial Transformer architecture.

Methods	TR1	TR2	OIO	OMP	CP	SDO	LDO	OR	Others	AVE
T w/ VF	62.66	60.72	63.07	42.52	69.18	63.71	61.63	64.05	63.68	62.86
RST w/ VF	67.50	68.37	67.31	55.18	71.26	68.67	63.99	61.84	59.68	64.32
T w/ CL	51.39	58.28	61.90	41.53	62.76	54.80	60.49	56.09	62.38	57.63
RST w/ CL	54.89	62.30	60.49	47.03	61.47	58.25	53.00	58.28	61.76	58.46

object class label by a [MASK] token in the input, and the model is trained to predict the class label of the masked region. We use the cross-entropy loss for training and testing; the loss is the atypicality signal. Since the input of class labels are discrete textual tokens, we project them through an embedding layer before feeding to the transformer; at the output, we project the last hidden layer of the masked input back to the class label by a decoder which shares the same weight as the embedding layer. We follow the same experimental setting as with the visual features.

Results. Experimental results are shown in Table 11. We find that checking semantic compatibility (CL) is not as effective as checking the visual compatibility (VF) under the unsupervised setting. Thus, visual features contain more useful information, and only checking the semantic compatibility is not enough for solving this task. An intuitive reason is that the visual features of an atypical object and a typical object are different; however, the class label input does not have this information when the atypical object is correctly detected by Faster R-CNN. On the other hand, the object detector might find difficult to recognize objects in some atypical images. As a result, using the class labels as input features bring noise to the model when learning the contextual compatibility. We examine this conjecture by using the human annotation for describing the atypical objects (*i.e.* OBJ1 and OBJ2) in the ad images. An ideal object detector should recognize the image as either OBJ1 or OBJ2. Therefore, we examine the recall rate that OBJ1 or OBJ2 is among the top 36 predictions by the Faster R-CNN detector. Results in each atypical category are shown in Table 12.

Table 12: Object detector performance on recognizing the annotated objects for each atypicality category.

Recall	TR1	TR2	OIO	OMP	CP	SDO	LDO	OR	Others	AVE
OBJ1	39.35	26.27	14.39	40.00	17.02	37.14	22.78	21.67	6.55	22.25
OBJ2	17.86	20.93	17.30	/	/	13.63	/	22.89	/	19.72

The micro average of the recall is only around 20%, which verify our conjecture that atypical objects are hard to be recognized. We find that the object detector tends to recognize the atypical object by its shape more than its texture (recall on OBJ1 is higher than OBJ2 for TR1 and TR2). OMP and SDO are relatively easier to be recognized. However, object detector has the worst accuracy on OIO, CP and Others, which may explain why RST has a lower performance than T when using class labels (CL) as input features for these categories.

4.3.5 Limitations

While our study on detecting atypicality in persuasive imagery provides valuable insights and promising results, it is important to acknowledge certain limitations that should be considered in the interpretation of our findings.

First, the study relies on a limited dataset for training and evaluation purposes. The dataset of visual advertisements, may not fully capture the diversity and complexity of real-world visual media, which could introduce biases or limitations in the detection and interpretation of atypicality. Different cultures, contexts, and domains may utilize unique and culturally specific atypical representations that are not adequately explored in this study. Additionally, the dataset may not account for the evolving nature of persuasive techniques and may become outdated over time.

Second, our model relies on self-supervised learning and the extraction of visual features to detect atypicality. While our Relative-Spatial Transformer (RST) shows improved performance compared to standard baselines and other transformer architectures, there are still cases where both our models and the baseline fail to accurately classify images. For exam-

ple, our models struggle to capture metaphoric similarity and interpret symbolic meanings, indicating that more nuanced visual features and domain-specific knowledge are required in these scenarios. Further research should investigate ways to enhance the models' ability to handle such complex and subtle forms of atypicality.

Lastly, our model is not designed for predicting fine-grained atypicality categories, such as texture replacement, deformation, and other specific types of visual transformations. While our approach demonstrates effectiveness in detecting overall atypicality in persuasive imagery, it lacks the capability to distinguish and classify the specific nature of atypical transformations in a granular manner. The categorization of atypicality into more specific types can provide valuable insights into the underlying visual strategies employed in persuasive media and further enhance our understanding of the rhetorical techniques used.

4.4 Chapter Summary

In this chapter, we have demonstrated the effectiveness of modeling contextual compatibility as a self-supervised approach to detect atypicality in persuasive imagery. Furthermore, analyses by atypicality categories have shown that our developed Relative-Spatial Transformer especially improves the detection performance on atypicality transformations involving spatial interactions between objects. These experimental results strongly support our hypothesis **H2** that modeling contextual compatibility through self-supervised learning methodologies enables the detection of atypical images, with spatial interactions between objects being a key factor in the process. Finally, we have found that learning semantic compatibility by predicted class labels is not sufficient and visual features are essential for detecting atypicality.

5.0 Interpreting Symbolism

5.1 Introduction

In this chapter, our objective is to develop computational models that can effectively interpret the symbolism employed in persuasive images. Symbolism serves as a powerful rhetorical device in the media, enabling messages to be conveyed more persuasively and efficiently [125]. Decoding symbolism involves recognizing that a particular item (*e.g.*, a baby) is a stand-in for something else (*e.g.*, innocence). This ability to decode symbolism has wide-ranging applications, including understanding persuasive texts and visual media [74, 35, 1]. For instance, a social media moderator needs to identify seemingly innocuous phrases or objects that may indicate prohibited behavior. An intelligent writing tutor should be able to recognize appropriate or inappropriate usage of symbolism in student essays. Furthermore, a persuasive text/image generator can enhance message delivery by employing symbolism effectively.

Automatic understanding of symbolism is crucial for developing computational intelligence capable of making inferences about the implied meaning within media. However, decoding symbolism is a challenging task, even for humans. One significant hurdle is the difficulty in automatically acquiring knowledge about symbolic relationships. Symbols serve diverse purposes, ranging from representing figures of speech and modes of thought to denoting signs, passwords, and customs [53]. Some symbols may have a semantic interpretation, such as meronyms or hyponyms, while others may be culturally dependent or require complex reasoning chains. Consequently, some relationships are relatively easier to identify when the related items appear together, either within the same sentence or image. In contrast, other relationships are more challenging to discern when the represented item does not co-occur with the symbol itself. Moreover, symbolic relationships can be situational, where the same symbol may represent different concepts in different scenarios. For example, while a baby often symbolizes innocence, it can signify burden and responsibility when depicted in the arms of a harried parent.

Recent research suggests that language models contain factual and commonsense knowledge that can be extracted through fine-tuning or probing techniques [93, 126, 44, 26]. Additionally, multi-modality models like CLIP have demonstrated the acquisition of substantial knowledge through self-supervised pre-training. This raises the possibility that these models might also encapsulate implicit and abstract knowledge, making them powerful resources for facilitating reasoning and other AI applications. However, the extent to which these models capture symbolism-related knowledge remains unclear. To investigate this question, we introduce a new probe called *SymbA*. In Section 5.2, we delve into the challenges of creating such an evaluative framework, describe the constructed evaluative datasets, and present analytical tools for analyzing various types of symbolic relationships. To effectively evaluate the decoding of symbolism, it is crucial to isolate this aspect from potential visual recognition problems that may affect the performance of models. While some models may possess the capability to decode symbolism, the performance could be hindered by insufficient object recognition. To address this concern, our evaluative framework primarily focuses on the textual modality. By emphasizing the linguistic aspects of symbolism, we can assess models’ understanding and interpretation of symbolic relationships without being influenced by visual cues.

5.2 Analysis Probe Construction

We present the **SymbA** (**S**ymbolism **A**nalysis) probe as a methodology to evaluate models’ proficiency in decoding symbolism. SymbA comprises 1066 symbolic pairs, such as ”red - passion”. In these pairs, the former, typically a physical object or content, is referred to as the *signifier*, while the latter, typically a more conceptual symbolic reference, is called the *signified* [139]. Our symbolic pairs are derived from two datasets. One set consists of *conventional symbol pairs* compiled from commonly used symbols in English literature, which tend to be context-invariant. The other set is a subset sampled from a previous visual advertisement corpus [50], containing *situated symbol pairs* where humans have annotated the local context surrounding the signifier and the intended signified. By modifying the

prompt to include or exclude the local description, we can investigate the impact of situated context on symbolism decoding.

To analyze the evaluative outcomes, we propose two tools. Firstly, we employ a heuristic metric based on point-wise mutual information to quantify the semantic relatedness between the signifier and its signified. This metric allows us to differentiate between “easier” pairs and “harder” pairs, providing insights into the difficulty levels of symbolic relationships. Secondly, drawing inspiration from previous work on commonsense relationships [119], we define a taxonomy of symbolic relationships (*e.g.*, UsedFor, HasProperty, etc.). By categorizing the symbolic pairs into these relationships, we can conduct fine-grained analyses to identify which types of symbolic relationships pose greater challenges for the models. Together, these tools enable us to gain a deeper understanding of the models’ performance in decoding different symbolic relationships.

5.2.1 Symbolism Data Sources

Conventional Literary Symbolism Based on the sheer volume of pretraining text a language model has seen, it seems plausible that the language model should have come across the more conventional, widely-used symbols. In these cases, the signified might almost be seen as an additional word sense for the signifier. Such symbolic relationships are often taught in high-school English classes as well as various writing courses and online resources.

Consulting multiple sources [12, 39, 119], we created a dataset of conventional symbolism that consists of 132 signifiers that are commonly used in literature. Our dataset covers a diverse set of signifiers that can be categorized into 11 groups¹, as shown in Table 13. Object, Animal, Plants and Nature are the most frequent types; while Action, Directions, Number and Christian has limited instances in the dataset. There are a total of 536 signifier-signified pairs, as each signifier may have multiple signifieds. The vocabulary size of possible signifieds is 333.

Situated Symbolism Situated symbolism refers to symbols that arise from specific circumstances and are not established by conventions. There is a great deal of variation

¹We consulted an educational website <https://www.dvUSD.org/cms/lib/AZ01901092/Centricity/Domain/2891/Gawain%20Symbols.pdf>

Table 13: Signifier types of conventional literary symbolism.

Signifier Type	Count	Example (signifier: signified)
Color	12	pink: femininity, flesh, ...
Nature	17	dawn: hope, illumination
Plants	18	rose: beauty, love, ...
Weather	9	mist: confusion, mystery, ...
Animal	19	lion: pride, power, ...
Setting	14	forest: evil, mystery, ...
Object	22	trophy: victory
Action	3	kiss: intimacy, fellowship, ...
Number	7	seven: creation, abundance, ...
Christian	7	angel: messenger, purity, ...
Directions	4	west: descending, old

in terms of the challenge of the task. At an extreme, one might consider a literary author taking chapters to develop and evolve a symbol, such as the meaning of Hester Prynne’s “A” in “The Scarlet Letter”; such a grand scale is out of the scope of this work. Here, we focus on a more manageable context range, limited to the message conveyed in a static visual advertisement [50]. We chose this domain because the ad offers a self-contained narrative for the context; any symbolic reference has to either be resolved through information directly presented in the ad or relies on commonly shared knowledge by the viewers.

The advertisement dataset includes bounding boxes around the signifiers in each ad image and their corresponding signified symbol references (*e.g.* fitness, happiness, danger). Although the bounding boxes are provided, no textual annotation describes the signifier, as discussed in Section 2.2.2. Therefore, we supplemented the dataset with additional annotations.² The signified vocabulary size is 53. We opted to create a balanced dataset for evaluation by randomly sampling 10 ads from each signified group, resulting in a total of 530 instances.³ Each instance was randomly assigned to one of the 11 annotators, including the author, 2 collaborators who were familiar with the Ads dataset, and 8 Ph.D. students who were volunteers for the annotation. Given an ad image with a visual signifier in the bounding

²We considered a captioning generation model, training an attention-based captioning model [6] on the COCO datasets; however, the domain gap between symbolic and general non-ad image was too large for the resulting captions to prompt language models.

³We manually checked each instance for making sure there is no offensive content.



Figure 18: A situated symbolism sample. Each sample contains a signifier-signified pair and a localized description. Here the signifier is *sandal*; the signified is *freedom*; the localized description is *sandals that look like a butterfly*.

box and its ground-truth symbolic reference (*i.e.* signified), the annotator was asked to write a description that should be in a short noun phrase and capable of conveying its symbolic meaning, which we refer to as *localized description*. The annotation instruction can be found in the Appendix B. To evaluate the reliability of the annotated descriptions, we qualitatively checked the inter-rater agreement between 3 annotators for 20 samples. While their descriptions did not always use the exact same wording, we found that their descriptions expressed the same meaning 90% of the time. The head noun of each description was then manually annotated as the signifier (referred as a task *without context*), while the description itself served as the *context* for the signifier (cf. Fig 18, *sandal* is selected as the signifier, while *that look like a butterfly* is a context stimulus).

Human Evaluation To assess the reliability of the datasets, inter-rater agreement between human annotators was computed. For the computational models, the task involved selecting the correct signified from a fixed set of options (333 for conventional symbols and 53 for ad symbols). However, the same task may be challenging for a human. An alternative approach is to simplify the experiment by presenting four answer options, including randomly selected negative candidates from a fixed vocabulary, and selecting the correct answer from among these options. We sampled a total of 50 context-dependent instances from the

Table 14: Human evaluation for decoding symbolism.

	Conventional Symbolism	Advertising Symbolism	
		w/o context	w/ context
Raw agreement score	0.73	0.68	0.70
Krippendorff’s alpha score	0.64	0.57	0.60
Accuracy	0.77	0.71	0.68

situated ad set, as well as 50 context-independent instances. Among the context-independent instances, 22 were sampled from the conventional set, while the remaining 28 were obtained from the advertising set. We conducted the annotation of these 100 instances with the help of 8 Ph.D. students from diverse cultural backgrounds (*e.g.*, European, American, and Chinese). To ensure impartiality and avoid any bias, both the author and collaborators refrained from participating in the annotation task, and the annotators were deliberately kept uninformed about any prior knowledge related to the dataset. Each instance was annotated by two human annotators. Given a signifier in textual modality, human annotators were asked to select the most appropriate signified from four answer options. The raw and adjusted inter-rater agreement scores are shown in Table 14, indicating moderate to substantial agreement [60, 40]. This demonstrates the quality of our data. Human performance on these tasks is also reported in Table 14, highlighting the challenges even for humans. It is important to note that the human annotators represent a variety of cultural backgrounds and have not received task-specific training, thus reflecting the ability of a typical person rather than the expertise of literary experts.

5.2.2 Probing Methods

In this work, we employ the method of using cloze statements as prompt templates to probe knowledge from language models [93]. By formulating prompts with missing tokens, we prompt the models to predict the exact missing token based on the provided context. This approach is computationally efficient and does not require specific preparation of negative instances. While other methods such as mining, paraphrasing [52, 41, 26] or implicit prompt-

ing [69, 99] have been explored to discover better prompts, we focus on the unsupervised elicitation of knowledge and do not involve training prompt-engineering methods.

To probe knowledge from multi-modality models, we directly feed the symbolic image into the image encoder. Specifically, for the CLIP model, we compute the cosine similarity score between the image and text embedding when replacing the missing token in the textual prompt with each signified candidate. The candidate with the highest score is considered as the top prediction.

5.2.3 Analytical Tools

Semantic Relatedness For quantitatively measuring the semantic relatedness between the symbolic pair, we develop a heuristic metric based on the pointwise mutual information. This metric measures how frequently a signifier-signified pair co-occur within the same sentences in a textual corpus. We assume that if the pair co-occur frequently, then the symbolic relationship leans towards a factoid thus considered as “easy” knowledge; on the other side, if the pair rarely co-occur in the same sentence, then it leans towards implicit commonsense reasoning thus considered as “hard” knowledge. So we use this metric for measuring the knowledge difficulty.

For a given signifier x and signified y , the PMI score is computed by

$$pmi(x, y) = \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{\frac{N(x, y)}{N}}{\frac{N(x)}{N} \frac{N(y)}{N}}$$

where $N(x, y)$ is the number of sentences containing both x and y ; $N(x)$ or $N(y)$ is respectively the number of sentences containing x or y ; N is the total number of sentences in the corpus.

A higher PMI score signifies a stronger and more easily recognizable symbolic relationship between two words. For example, the PMI score of 4.61 between “ornaments” and “Christmas” in the BookCorpus dataset [150] indicates a highly semantically related symbolic association. Conversely, a lower PMI score indicates a more challenging or distantly-related symbolic connection between words. For instance, the pair “apple” and “sin” has a

Table 15: Relationship type distribution of *signifier-signified* in the set of advertising symbolism.

Relationship Type	Count	Example (signifier - signified)
UsedFor	52	makeup - beauty
HasProperty	46	child - youth
RelatedTo	47	mountain - adventure
Others	94	chocolate - love (SymbolOf)
Indirect	116	giraffe - love

PMI score of -3.20, indicating a less obvious or distant symbolic association between these terms.

Symbolic Relationship Types For investigating the fine-grained types of each symbolic relationship, we further annotate each signifier-signified pair according to a pre-defined taxonomy of commonsense relationships [119]. The symbolical associations used in ads are creative and diverse, while the conventional set mostly contains the narrowly-defined symbolic relationship (*i.e.*, SymbolOf in ConceptNet [119]). Therefore we conduct this analysis on the ads set only. As shown in Table 15, we specifically study the three most frequent types (*i.e.*, UserFor, HasProperty, and RelatedTo) that appear in the advertisement set. Minor types, such as Synonym, Antonym, IsA, Causes, SymbolOf, etc., are combined into a category named Others. Symbolism knowledge that cannot be clearly determined is classified as Indirect.

5.3 Re-ranking Approach for Bias Mitigation

The bias in language models’ pre-training corpus may have negative impact on knowledge elicitation. A model’s prediction candidates that appear more frequently in the pre-training corpus tend to be ranked higher than its appropriate position; similarly, rarer signifieds may be unfairly penalized. For example, the language model may consider “freedom” as a more probably predicted candidate than “serenity” since the latter word has been rarely

seen during the pre-training. In order to reduce the bias effect brought by the pre-training frequency, we propose a new approach for ranking the predictions by considering the prior probability of each candidate.

Assuming that x represents the signifier, y represents the signified, t represents the prompt (e.g. “is a symbol of”) and θ represents the parameters of the language model, the conditional probability of y is represented as $p(y|x, t, \theta)$. In the common way, the top candidate y_{pred} is selected by having the highest probability: $y_{pred} = \operatorname{argmax}_y p(y|x, t, \theta)$ [93, 52]. In our approach, we re-rank the previously-selected top k candidates after normalizing the conditional probability by the prior probability of each candidate:

$$y_{pred}(k) = \operatorname{argmax}_{y \in Y_k} \log \frac{p(y|x, t, \theta)}{p(y|t, \theta)}$$

where Y_k is the set of previously-selected top k candidates. The intuition is that a high $p(y|x, t, \theta)$ might not mean a good collocation between x and y if $p(y|t, \theta)$ is also high. For example, a certain signified (e.g. love) might have a high probability when following the prompt (e.g. “is a symbol of”), no matter which signifier is given. Our re-ranking approach aims to reduce this bias effect.

5.4 Experiments

In this section, we present the experiments conducted to evaluate the performance of different language models and multi-modality models in decoding symbolism. We also investigate the bias problem and measure the effectiveness of the debiasing method. Additionally, we analyze the fine-grained performance of the models based on knowledge difficulty and relationship types.

5.4.1 Setup

We compare five language models that represent different pre-training strategies, architectures and sizes: Word2Vec [78], BERT [28], RoBERTa [73], GPT-2 [103] and GPT-J-6B

[134]; and a representative multi-modality model: CLIP [100]. As for baseline models, we consider random guessing and co-occurrence ratio.

Random Baseline: rank signified candidates by a random order (average over 10 random runs).

Co-occurrence Baseline: rank signified candidates by its co-occurrence ratio with the signifier according to BookCorpus [151]. The ratio is computed by $\frac{N(x,y)}{N(y)}$ with the same notations as defined in Section 5.2.3.

Word2Vec: rank signified candidates by the cosine similarity between the signifier word vector and each signified candidate vector. For situated symbolism, the signifier word vector is replaced by the context vector that is the summation of each token vector in the localized description.⁴

BERT (336M parameters): rank signified candidates by the probability of the masked token by querying the language model with a cloze prompt (i.e. “[signifier] is a symbol of [MASK].”)⁵. For decoding general symbolism, “[signifier]” is replaced by the signifier token; for decoding situated symbolism, “[signifier]” is replaced by the localized description of the signifier.⁶ Notice that the majority of signifieds are tokenized as single word pieces, with only around 20% requiring multiple word pieces. For these cases, we use the stemmed piece to transform them into a single word piece.

RoBERTa (355M parameters): same as BERT.⁷

GPT-2 (124M parameters): rank signified candidates by the probability of the next token by querying the language model with the first part of the sentence (i.e. “[signifier] is a symbol of”).⁸

GPT-J (6B parameters): same as GPT-2.⁹

CLIP (152M parameters): rank the signified candidates by calculating the cosine similarity score between the image and text embeddings. We replace the masked token in the

⁴‘word2vec-google-news-300’ in gensim 4.1.2

⁵Since prompt selection is not a focus of this work, we simply picked a prompt that echoes the surface text for the “SymbolOf” relation presented in ConceptNet [119].

⁶‘bert-large-uncased’ in transformers 4.8.2

⁷‘roberta-large’ in transformers 4.24.0

⁸‘gpt2’ in transformers 4.8.2

⁹‘EleutherAI/gpt-j-6B’ in transformers 4.24.0

Table 16: Model performance (P@n) for decoding symbolism.

	Conventional Symbolism			Advertising Symbolism					
	P@1	P@5	P@10	w/o context			w/ context		
				P@1	P@5	P@10	P@1	P@5	P@10
Random	1.29	5.15	10.45	2.48	11.43	23.83	2.12	9.77	20.30
Co-occur	7.58	18.94	35.61	16.10	42.86	57.89	13.96	34.53	46.42
Word2Vec	5.30	25.76	46.21	18.42	43.23	57.89	14.53	32.64	47.17
BERT	10.61	27.27	40.15	10.15	25.56	39.85	11.51	27.17	39.81
RoBERTa	19.70	33.33	42.42	13.16	33.08	45.86	10.00	27.55	45.47
GPT-2	6.06	16.67	26.52	4.51	17.67	30.08	7.36	19.43	37.74
GPT-J	27.27	46.97	56.06	10.90	28.20	42.48	13.96	33.77	50.00
CLIP	/	/	/	/	/	/	21.13	48.30	63.02
GPT-J (open vocab)	15.15	39.39	48.48	2.63	11.28	16.92	4.91	13.02	18.68

textual prompt with each signified candidate (*i.e.*, “a symbol of [MASK]”).¹⁰ Since the multi-modality model requires both visual and textual input, we only evaluate its performance on the set of advertising symbols with context. We use the corresponding ad image as the visual input.

We evaluate each model based on how highly it ranks the ground-truth signified against others in a fixed vocabulary. We also evaluate GPT-J’s performance under an open-vocabulary setting. We use the precision at n ($P@n$) as the evaluative metric. To account for multiple valid signifieds for a given signifier, this value is 1 if at least one of the valid signifieds is ranked among the top n predictions, and 0 otherwise. Experiments are conducted on the GPU model of NVIDIA Quadro RTX 5000, 16G memory, driver version 460.84 and CUDA version 11.2.

5.4.2 Model Performance on Decoding Symbolism

In the experiments evaluating the performance of different language models and CLIP on decoding symbolism, several key findings emerge.

Newer LMs outperform their previous iterations. Table 16 shows the overall performance for decoding symbolism through our SymbA probe. Newer language models, such as GPT-J, outperform their previous iterations. GPT-J shows the best overall per-

¹⁰ViT-B/32’ in OpenAI’s clip API

Table 17: Model performance (P@1) on each signifier group of conventional literary symbolism.

	Color	Nature	Plants	Weat.	Anim.	Setting	Object	Action	Num.	Christ.	Direct.
RoBERTa	50.00	35.29	11.11	11.11	10.53	7.14	31.82	0.00	0.00	14.29	0.00
GPT-J	41.67	35.29	33.33	33.33	36.84	7.14	27.27	33.33	0.00	14.29	0.00

formance for decoding conventional symbols, surpassing other models. Even in the more challenging open-vocabulary setting, GPT-J performs comparably to BERT or RoBERTa in the fixed-vocabulary setting. Scaling up the same type of language model leads to substantial improvements, with GPT-J performing 21 points better than GPT-2 and RoBERTa performing 9 points better than BERT in $P@1$.

Variations in signifiers’ types impact decoding. Table 17 compares RoBERTa and GPT-J’s performances by signifier types. Both models excel at decoding *Colors* but falter on *Numbers* and *Directions*. On average, GPT-J outperforms RoBERTa, although it has lower accuracy for *Colors* and *Objects*. We conjecture that the Web data used to pre-train GPT-J may be more multi-modal such that color attributes may be shown visually.

CLIP has a superior performance on situated ad symbols. As shown in Table 16, CLIP demonstrates significantly better performance than language models when decoding situated ad symbols. Among LMs, Word2Vec has the best $P@1$, and GPT-2 has the worst. It is surprising that powerful language models such as RoBERTa perform worse than the simple Word2Vec or the Co-occur baseline on this task. We have similar observations for decoding the advertising symbolism without context. The prior-bias problem faced by advanced language models may contribute to their decreased performance in decoding symbolism, while CLIP’s multi-modal nature allows it to better associate visual and textual information.

5.4.3 Effectiveness of Debiasing

The data bias problem exists, and re-ranking significantly reduces it. We first compute the correlation between each signified’s (y_i) frequency and its predicted probability, $p(y_i|x, t, \theta)$ for verifying the existence of the bias introduced in Section 5.3. We use

Table 18: Pearson correlation scores between candidates’ frequency and prediction probability before or after normalized by the prior probability.

Model	Pearson score before	Pearson score after
BERT	0.375	-0.107
RoBERTa	0.355	-0.123
GPT-2	0.483	-0.192
GPT-J	0.363	-0.244

BookCorpus [150] as the source for estimating y_i ’s frequency and use the advertising symbolism as testing samples. The Pearson correlation scores are reported in Table 18. Initially, the Pearson scores are all above 0.3, considered to be positively moderate, indicating the presence of bias [24]. However, after applying the re-ranking approach that considers the prior probability of the signified¹¹, the correlation decreases to a low level, from -0.107 to -0.244, which can be interpreted as no or slight correlation [24], suggesting a mitigation of the bias. Although the absolute correlation score decreases, there is a shift from a positive to a negative correlation level, indicating an over-correction of the bias.

Debiased LMs and CLIP rival human performances in some cases. As shown in Table 19, language models after re-ranking have better performance on decoding symbolism than the original ones. In particular, the improvement for larger models such as RoBERTa is more than 200% on decoding ad symbolism. The re-ranking approach boosts RoBERTa to a relatively high accuracy, 25.19 (or 26.04) for decoding ad symbolism without (or with) the situated context. Debiased RoBERTa or GPT-J surpasses CLIP’s performance in decoding situated symbols. This suggests that LMs are effective tools for decoding symbolism when visual content can be properly translated into textual descriptions. We conducted a simplified 4-choice task on the same dataset used for assessing human performance to further evaluate these models. Surprisingly, the results indicate that debiased GPT-J outperforms humans in understanding conventional symbolism, as presented in Table 20. For ad symbols, debiased RoBERTa achieves performance close to that of humans, with only a 4-point difference.

¹¹By considering the prior probability of y_i , we compute the Pearson correlation score between y_i ’s frequency and $\frac{p(y_i|x,t,\theta)}{p(y_i|t,\theta)}$.

Table 19: Measuring the effectiveness (P@1) of the re-ranking approach for decoding symbolism (original \rightarrow re-ranked).

	Conventional	Advertising	
		w/o context	w/ context
BERT \rightarrow_R	10.61 \rightarrow 12.88	10.15 \rightarrow 17.29	11.51 \rightarrow 22.08
RoBERTa \rightarrow_R	19.70 \rightarrow 20.45	13.16 \rightarrow 25.19	10.00 \rightarrow 26.04
GPT-2 \rightarrow_R	6.06 \rightarrow 7.58	4.51 \rightarrow 9.77	7.36 \rightarrow 19.43
GPT-J \rightarrow_R	27.27 \rightarrow 28.03	10.90 \rightarrow 22.18	13.96 \rightarrow 22.82

Table 20: Accuracy on the multi-choice task: human versus LMs (original \rightarrow re-ranked) and CLIP.

	Conventional	Advertising	
		w/o context	w/ context
Human	77.27	71.43	68.00
RoBERTa \rightarrow_R	68.18 \rightarrow 77.27	35.71 \rightarrow 67.86	42.00 \rightarrow 64.00
GPT-J \rightarrow_R	72.73 \rightarrow 90.91	53.57 \rightarrow 64.29	50.00 \rightarrow 62.00
CLIP	/	/	60.00

Debiased RoBERTa and GPT-J demonstrate different strengths. Table 19 and Table 20 show that GPT-J performs better in decoding conventional symbols, while RoBERTa excels in decoding advertising symbols. Further analysis is conducted to explain these observations in the subsequent section.

5.4.4 Fine-grained Performance with Analytical Tools

The fine-grained analysis using the analytical tools in the SymbA probe provides insights into the situations where LMs fail and how the re-ranking approach helps improve their performance. With the fine-grained relationship types, we analyze LMs and CLIP’s behaviors in decoding different types of symbolism.

Analysis by Knowledge Difficulties: 1) RoBERTa performs better on semantically-related symbols while GPT-J excels in distantly-related symbols. We first measure the difficulty distribution of both symbolism sets. The knowledge difficulty of symbolic pairs is measured using the PMI score, as explained in Section 5.2.3. The mean of PMI scores for the ad set and the conventional set are respectively -0.997 (with ± 1.56 variance) and -3.872 (with ± 5.96 variance). The PMI distribution of both sets is shown in Figure 19. It indicates that the symbolism samples in the ad set are generally easier than those in the conventional set. To gain more insights, the samples are split into different difficulty groups based on their PMI scores. The model performance for each difficulty group is reported in Table 21. The conventional set contains mostly hard cases (only 5% of them have $\text{PMI} > -2$). The knowledge difficulty of ads symbolism is more diverse, covering both easy and hard ones. It is observed that GPT-J performs better on harder cases and struggles with easier cases. GPT-J_R, in particular, performs well when the PMI score is extremely low, suggesting its ability to interpret very rare symbols.

2) Debiasing improves the performance of semantically-related symbolic pairs without significantly affecting distantly-related ones. By comparing the model performance before or after re-ranking in Table 21, we find that the re-ranking approach leads to substantial improvements for both RoBERTa and GPT-J in decoding easy cases (up to 62% increase on $P@1$ for $\text{PMI} > 1$), with less decrease in performance for hard cases. The intu-

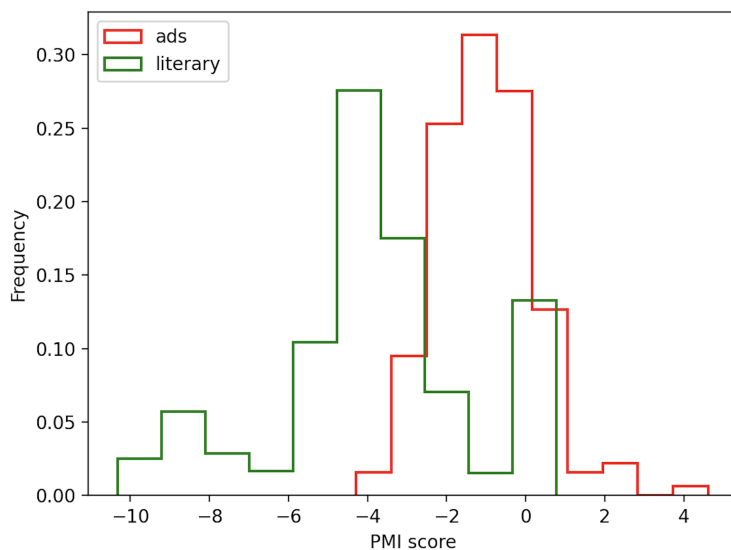


Figure 19: Knowledge difficulty distribution in the conventional (green) and the advertising (red) symbolism.

Table 21: Model performance ($P@1$) on the conventional literary symbolism (upper) and the advertising symbolism (lower), on different PMI scores (measure of difficulty, from high to low). Comparing RoBERTa with GPT-J, the higher $P@1$ is bolded. Comparing the effectiveness of the re-ranking approach (original \rightarrow re-ranked), the improvement is marked in green and the drop is marked in red. We also provide an example in each PMI group for gaining more insights.

PMI score	-inf (75)	<-6 (76)	-6 to -5 (37)	-5 to -4 (136)	-4 to -3 (129)	-3 to -2 (56)	>-2 (27)
(Example)	blue - conservatism	gold - dominion	ladder - connection	night - death	apple - sin	dove - purity	three - tripartite)
RoBERTa \rightarrow_R	1.33 \rightarrow 1.33	5.26 \rightarrow 5.26	5.41 \rightarrow 0.00	5.88 \rightarrow 0.74	6.20 \rightarrow 8.53	3.57 \rightarrow 8.93	3.70 \rightarrow 18.52
GPT-J \rightarrow_R	1.33 \rightarrow 4.00	7.89 \rightarrow 2.63	5.41 \rightarrow 2.70	7.35 \rightarrow 4.41	6.98 \rightarrow 6.98	5.36 \rightarrow 16.07	18.52 \rightarrow 22.22
PMI score	-inf (20)	<-2 (79)	-2 to -1 (108)	-1 to 0 (87)	0 to 1 (45)	>1 (16)	
(Example)	igloo - refreshing	gun - death	bird - freedom	dragon - adventure	beach - vacation	ornaments - christmas)	
RoBERTa \rightarrow_R	5.00 \rightarrow 5.00	6.33 \rightarrow 5.06	12.04 \rightarrow 10.19	10.34 \rightarrow 18.39	13.33 \rightarrow 48.89	6.25 \rightarrow 68.75	
GPT-J \rightarrow_R	5.00 \rightarrow 10.00	6.33 \rightarrow 1.27	10.19 \rightarrow 7.41	8.05 \rightarrow 17.24	8.89 \rightarrow 51.11	6.25 \rightarrow 50.00	

Table 22: The PMI score for each relationship type.

Relationship Type	PMI mean \pm variance
UsedFor	-0.39 \pm 2.35
HasProperty	-1.02 \pm 1.31
RelatedTo	-0.86 \pm 0.75
Others	-0.51 \pm 1.33
Indirect	-1.71 \pm 0.93

ition behind this improvement is that the prior probability of the signified, as a denominator term for computing the PMI score, tends to be small when PMI is large (*i.e.* easy cases). So normalizing by this small prior probability increases the ranking of the correct signified for easy cases. Similarly, the performance on hard cases after re-ranking is expected to decrease. Interestingly, the impact of the re-ranking approach is considerably positive for easy cases and only slightly negative for hard cases, resulting in an overall improvement. Examining their performance in different difficulty groups, it is observed that the accuracy of GPT-J_R and RoBERTa_R generally increases as the knowledge difficulty decreases. Surprisingly, the original models exhibit relatively stable performance and even perform slightly worse on the easiest cases (PMI > 1).

Analysis by Relationship Types: 1) Breakdown by relationship types is consistent with analysis by knowledge difficulties. We first measure the difficulty level of each relationship type introduced in Table 15. We show the result in Table 22. Indirect is identified as the most difficult type (because the logical reasoning between these symbolic pairs is hard to identify), while UserFor is the easiest. The model performance for each relationship type is presented in Table 23. Consistent with previous observations, the re-ranking approach improves the decoding accuracy more for the types of *UsedFor*, *Others* and *RelatedTo*, which are relatively easier (PMI > -1) compared to other types. Moreover, RoBERTa outperforms GPT-J in decoding these types of symbols.

2) Debiasing improves LMs’ robustness without prompt engineering. To further investigate the impact of prompt engineering, type-specific prompts are used for each relationship type (*e.g.*, the default “is a symbol of” is replaced by “is used for” when prob-

Table 23: Model performance ($P@1$) on relationship types when using the default prompt (“is a symbol of”) or a type-specific prompt (respectively “is used for”, “has the property of” or “relates to” for the relationship type of “UsedFor”, “HasProperty” or “RelatedTo”).

Relationship type	UsedFor		HasProperty		RelatedTo		Others	Indirect
	default	specific	default	specific	default	specific	default	default
RoBERTa	5.77	23.08	10.87	4.35	8.51	4.26	20.21	3.45
RoBERTa _R	21.15	21.15	15.22	17.39	19.15	14.89	37.23	4.31
GPT-J	9.62	19.23	10.87	19.57	4.26	2.13	14.89	2.59
GPT-J _R	21.15	23.08	17.39	26.09	17.02	10.64	28.72	3.45
CLIP	21.78	/	30.77	/	14.52	/	25.17	13.77

ing a symbol in the type of UsedFor). We find that the type-specific prompt can sometimes greatly facilitate the original models on decoding knowledge: RoBERTa increases 17 points for UsedFor; GPT-J increases around 9 points for UsedFor or HasProperty. At first glance, this suggests that these LMs do have knowledge about the semantic relationships between the signifier and signified, but the general prompt cannot elicit the desired response. However, it is observed that type-specific prompts have little impact on the re-ranked models, *e.g.*, RoBERTa performs same when prompted by the default or the type-specific template. The re-ranking approach helps stabilize the performance of LMs, indicating that improving debiasing methods is more crucial than prompt engineering for developing robust models.

3) CLIP exhibits different behaviors compared to LMs when decoding different types of symbolism. CLIP performs better in decoding the types of HasProperty and Indirect. This difference in performance may be attributed to the importance of visual content in understanding these types of symbolism. For example, for the HasProperty type, the model may require more detailed visual information about the signifier object to interpret the symbolic implication of the property. Similarly, for cases of Indirect, the visual patterns may contain denser and more useful information than the textual description, making the visual information crucial for interpreting these types of symbolism.

Table 24: Performance on atypical versus non-atypical advertising images.

	Atypical		Non-atypical	
	P@1	P@5	P@1	P@5
RoBERTa _R	19.61	47.06	18.18	54.55
CLIP	25.49	49.02	18.18	54.55

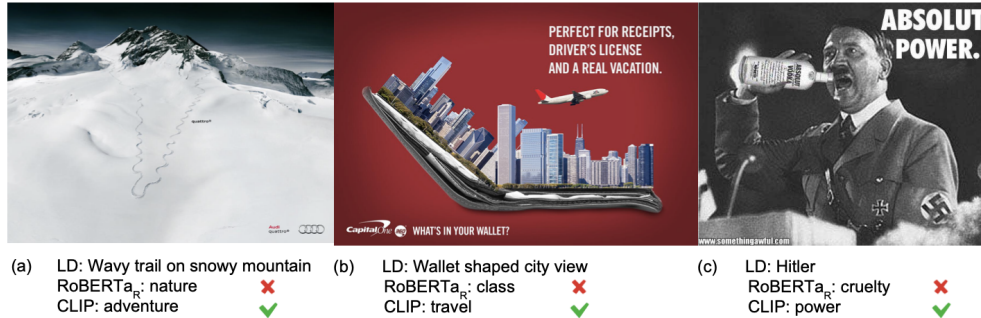


Figure 20: Case study for comparing predictions from RoBERTa and CLIP.

5.4.5 Performance in Atypical Images

In Chapter 4, we investigated the rhetoric of atypicality in advertising images. To explore the relationship between atypicality and symbolism interpretation in advertising images, we specifically evaluate the performance of models in decoding symbolism on both atypical and non-atypical images.

To accomplish this, we curate a subset of our situated symbolic set comprising 62 images, consisting of 51 atypical and 11 non-atypical images, based on the atypicality annotations [143]. Using the debiased RoBERTa and CLIP models, which demonstrated the best performance in our previous experiments, we predict the symbolic reference of these images. The RoBERTa model utilizes localized descriptions (LD) as input, while the CLIP model uses the images themselves as input.

The results, shown in Table 24, reveal interesting findings. We observe that RoBERTa and CLIP exhibit similar performance on non-atypical images. However, when it comes to decoding symbolism in atypical images, CLIP outperforms RoBERTa significantly. These

findings highlight the effectiveness of CLIP in decoding symbolism within atypical advertising images. The superior performance of CLIP suggests its ability to capture and interpret the nuanced and unconventional visual strategies employed in atypical imagery. Notably, Figure 20 provides illustrative examples where CLIP successfully decoded the symbolism, whereas RoBERTa failed. Hilter can be employed as a symbol representing either cruelty or power. However, in Figure 20c, the posture of Hilter, specifically holding Vodka like a microphone, signifies "power" as a more accurate and fitting reference. By leveraging the visual information directly, CLIP showcases its potential in understanding the symbolism embedded in persuasive atypical advertisements. This further emphasizes the importance of considering multimodal approaches, such as CLIP, for a comprehensive analysis of symbolism in persuasive visual media.

5.4.6 Limitations

Because decoding symbolism is a challenging new problem, our approach and experimental results have some limitations. Our work builds on available resources, which may have a bias toward an English/Euro-centric perspective. Additionally, the evaluative datasets that we curated have a limited coverage of possible symbols even within the English literary tradition. The symbolism datasets used may not fully capture the diversity and complexity of symbolism in various domains, leading to potential limitations in generalizability. Moreover, biases present in the data, such as cultural or regional biases, can impact the performance and interpretation of the models.

Secondly, as mentioned in Section 5.2.1, our study on situated symbolism is limited to symbolic pairs that can be found in static visual advertisements rather than longer form text or videos. However, symbolism can be highly context-dependent and influenced by real-time factors such as social, cultural, and personal contexts. The absence of real-time contextual information in our experiments may restrict the models' performance in understanding and interpreting symbolism accurately in dynamic and evolving situations.

Finally, although our study incorporates visual information through CLIP, the reliance on textual descriptions and prompts may limit the models' ability to fully leverage the visual

context. Symbolism often involves visual cues, and the textual representations alone may not capture the richness and subtleties present in visual symbolism. Further research exploring ways to enhance the models’ understanding and utilization of visual content could address this limitation.

5.5 Chapter Summary

In this chapter, we evaluated the feasibility of extracting symbolic knowledge from different language models and CLIP. Through the SymbA probe, we assessed their performance and achieved significant insights. We found that advanced large language models like GPT-J and RoBERTa, after undergoing debiasing, demonstrated human-level performance in identifying the intended signified concept from a given signifier. While CLIP’s overall performance in decoding situated symbols was slightly lower than language models, it exhibited specific strengths in certain types of symbolism, such as HasProperty and Indirect relationships. These results validated hypothesis **H3**, highlighting the potential of incorporating pre-trained models as a valuable source of knowledge for understanding and interpreting symbolism.

6.0 Conclusion

6.1 Summary

In this thesis, we have undertaken a comprehensive investigation into the modeling of visual rhetorics for persuasive media using self-supervised learning. Our research has focused on understanding and capturing the persuasive elements present in visual media, encompassing different facets of rhetorics. We have hypothesized that our goals could be achieved through self-supervised learning methods by harnessing general data without persuasion-related labels (Sec. 1.2). In Chapter 3, we have created a multi-modal dataset, specifically designed to analyze the persuasion modes exhibited in tweet images. We have employed novel annotation strategies to ensure the reliability of the annotated persuasion labels. Building upon this dataset, we have developed a self-supervised multi-modality model that was pre-trained on image-text pairs extracted from tweets. Our experimental results have provided support for our first hypothesis (**H1** in Sec. 1.2). In Chapter 4, we have presented a novel self-supervised approach for detecting persuasive atypicality in advertising images. Its competing performance has provided compelling evidence in support of our second hypothesis that atypical images can be detected by modeling contextual compatibility and spatial interactions between objects (**H2** in Sec. 1.2). In Chapter 5, we have constructed a novel evaluative framework designed to assess models' ability to interpret symbolism. Our objective is to investigate whether advanced large language models and CLIP, through simple self-supervised learning tasks, have acquired substantial knowledge of symbolism that can be utilized for interpreting the symbolic elements present in persuasive images. Empirical experiments conducted within our evaluative framework have provided strong support for our third hypothesis (**H3** in Sec. 1.2). Finally, we have explored the relationship between atypicality and symbolism interpretation in advertising images.

The following is a summary of our contribution:

- We have conducted the first study exploring the persuasion modes of images in social media. Our work has revealed the mutual influences between persuasion modes, per-

suasiveness, visual content, and political ideology, enhancing our understanding of how persuasion operates in multi-modal contexts.

- We have introduced a new multi-modal dataset, ImageArg, with annotations of social stance, image-enhanced persuasiveness, visual content, and modes of persuasion. This dataset has advanced multimodal persuasive media analysis.
- We have proposed a self-supervised multi-modality model for predicting persuasion modes. Our model has shown better performance in certain cases, encouraging future exploration in this direction.
- We have pioneered the study on unsupervised detection of persuasive atypical advertising images. Our research has opened up possibilities for refining unsupervised detection methods, considering additional cues, and incorporating larger and more diverse datasets.
- We have demonstrated the effectiveness of modeling visual compatibility in detecting atypical persuasive images. This self-supervised objective potentially has broader applications in identifying images that deviate from typical representations.
- We have proposed a novel technique for effectively modeling spatial interactions between objects. Our approach has offered opportunities for future research in improving performance and applicability across different domains and types of persuasive media.
- We have conducted the first comprehensive study assessing language models and multi-modality models in decoding symbolism. Our work has revealed their ability to learn implicit and abstract knowledge through self-supervised learning tasks.
- We have presented a new evaluative framework, SymbA, consisting of symbolic data, analytical tools and a debiasing method. Our thorough analysis has provided insights into model performance, especially their weakness that could be improved. SymbA can facilitate future explorations in evaluating the performance of new models in decoding symbolism, and set a standard for constructing new frameworks and methodologies in various tasks related to symbolic analysis.

6.2 Future Work

We consider the following future work that aim to build upon the foundations laid by our research and extend the knowledge and insights gained to advance the field of visual rhetorics and persuasive media analysis.

- **Expanding ImageArg dataset:** Extend the ImageArg dataset by including more diverse social topics and a larger number of annotated images. This will enhance the dataset’s representativeness and enable more comprehensive analysis of persuasive media in social contexts.
- **Refining and optimizing self-supervised models:** Further explore and refine the self-supervised multi-modality model for predicting persuasion modes. Investigate different architectural variations, training strategies, and data augmentation techniques to improve its performance and generalizability.
- **Enhancing unsupervised detection methods:** Continuously refine and enhance the unsupervised detection methods for identifying persuasive atypical advertising images. Consider incorporating additional visual and textual cues, such as image captions or metadata, to improve the accuracy and reliability of the detection process.
- **Advancing spatial interaction modeling:** Further develop and expand the proposed technique for modeling spatial interactions between objects in persuasive images. Explore its applicability in different domains and types of persuasive media, such as political campaigns, brand advertising, or social media content, to improve the detection and interpretation of persuasive visual cues.
- **Advancing symbolic analysis in language and multimodal models:** Continue exploring and refining the ability of language models and multimodal models to decode symbolism. Investigate the potential of more advanced models, *e.g.* ChatGPT and GPT-4 [86], novel self-supervised learning tasks, or fine-tuning techniques to enhance their understanding and interpretation of symbolic relationships in persuasive images.
- **Further research on bias mitigation:** Continue investigating and developing techniques to mitigate bias in language models and other computational models used in symbolic analysis. Explore additional debiasing methods and evaluate their effectiveness

in reducing bias and improving the performance of models in decoding symbolism and other related tasks.

- **Extending SymbA framework:** Extend the SymbA framework by incorporating additional types of symbolic data from other cultural backgrounds. Establish it as a comprehensive and versatile framework for evaluating models in various symbolic analysis tasks.
- **Real-world applications and impact:** Apply the findings and methodologies of our research to real-world applications in advertising, marketing, social media analysis, and content creation. Collaborate with industry partners to develop practical tools and systems that assist advertisers, marketers, and analysts in creating more impactful and persuasive campaigns while understanding and mitigating potential ethical concerns.

Appendix A Annotation Instruction for ImageArg

A.1 Stance

We setup different instructions for stance annotations on different topics since we would like to provide detailed instructions and examples for different topics separately.

A.1.1 Stance: Gun Control

We aim to study the topic and the stance of tweets. Given a tweet accompanying with an image, you need to answer the stance of the tweet towards a given topic, as depicted in Figure 21. Please make sure that you have the basic knowledge about that social topic and you understand the key message that the tweet (i.e. both the text and the image) sends. Just skip the HIT if you are not sure.

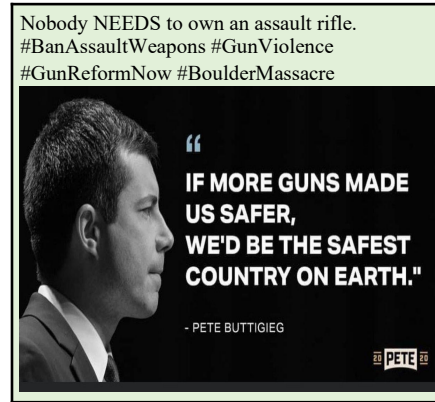
The question is about the stance. You need to decide whether the tweet is relevant or not to the social topic gun control. If it is relevant, then you need to annotate the stance: supports/opposes to/doesn't hold any stance.

A tweet is considered as relevant if it talks about anything that has to do with, but not limited to, the following issue categories: the Second Amendment, Gun control laws, etc. Tweets which contain the following hashtags are probably relevant to gun control: #NoBillNoBreak, #WearOrange, #EndGunViolence, #DisarmHate, #molonlabe, etc.

A tweet should be considered as irrelevant if it mentions a gun death event or a gun violence news, but the context is not necessarily about gun control.

Some examples for relevant tweets and their stance (we only show the text here, but you need to answer this question from both the text and image):

- *“Standing up for the second amendment and carrying a firearm for self defense.”* This tweet asks the audience to stand up for the 2nd amendment, which opposes to gun control;
- *“I don't understand why we can't ban assault weapons. We all know they are only used for*



This tweet _____ the topic "gun control".

- supports
- opposes to
- doesn't hold any stance to
- is not relevant to

Figure 21: Example of stance annotation on gun control.

hunting people. #PrayForOrlando #guncontrolplease.” This tweet talks about banning weapons and contains the hashtag “#guncontrolplease”, which supports gun control;

- *A common way to reduce violence in schools is to implement stronger security measures, such as surveillance cameras, security systems, campus guards and metal detectors. #violence #domesticviolence #gun #gunviolence #abuse #people #world #person #workplace.*” This tweet is relevant to the topic, but we are not sure about its stance.

Some examples for non-relevant tweets (we only show the text here, but you need to answer this question from both the text and image):

- *“Love will always conquer hate. #PrayForOrlando #OrlandoShooting.”* This tweet talks about gun violence but not about gun control;
- *“#Gunviolence has serious and lasting social and emotional impacts on those who directly and indirectly experience it.”* This tweet points out the impact of gun violence but not about gun control.

A.1.2 Stance: Immigration

We aim to study the topic and the stance of tweets. Given a tweet accompanying with an image, you need to answer the stance of the tweet towards a given topic, as depicted in Figure 22. Please make sure that you have the basic knowledge about that social topic and you understand the key message that the tweet (i.e. both the text and the image) sends. Just skip the HIT if you are not sure.

The question is about the stance. You need to decide whether the tweet is relevant or not to the social topic immigration. If it is relevant, then you need to annotate the stance: supports/opposes to/doesn't hold any stance.

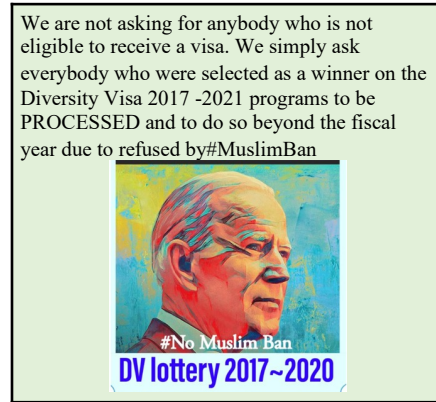
A tweet is considered as relevant if it talks about anything that has to do with, but not limited to, the following issue categories: Borders, Birthright citizenship, Immigrant Crime, DACA and the DREAM Act, Deportation debate, Economic impact, Immigration quotas, Immigrants' rights and access to services, Labor Market - American workers and employers, Law enforcement, Refugees, etc.

A tweet should be considered as irrelevant if it mentions a group of immigrant people such as Muslim, Syrian refugees but doesn't explicitly talk about immigration issues.

Some examples for relevant tweets and their stance (we only show the text here, but you need to answer this question from both the text and image):

- *“Man feels bad for new immigrant driver in Brampton that crashed into his truck, causing \$6K worth of damages - he had no licence or insurance”*. This tweet is related to the topic of immigration under the category of Immigrant Crime, and it opposes to immigration.
- *“House Bill 3438 will finally give our immigrant students some desperately needed resources! Thank you State Representative Maura Hirschauer for introducing this bill! Now, let's make sure this bill becomes law!”* This tweet is related to the topic of immigration under the category of DREAM Act, and it supports immigration.
- *“I'm a woman that supports Trump to fix economy, immigration, school, military more. #MAGA3X”* We consider a tweet as relevant even if it mentions several topics in addition to immigration, and it opposes to immigration.

Some examples for non-relevant tweets (we only show the text here, but you need to



This tweet _____ the topic "immigration".

- supports
- opposes to
- doesn't hold any stance to
- is not relevant to

Figure 22: Example of stance annotation on immigration.

answer this question from both the text and image):

- *“Will I die, miss?’ Terrified Syrian boy suffers suspected gas attack.”* This tweet talks about a Syrian boy suffering a gas attack, which may be pointing to a war or terrorist event in Syria, not necessarily directly about an immigration issue.
- *“Virtual tour of Steinbach, in partnership with MANSO, Welcome Place, Eastman Immigrant Services and the Steinbach LIP, coming up March 9th, 2021. It’s free so don’t miss out!”* This tweet mentions Immigrant Services, but does not talk about any immigration issue.
- *“I called on [USERNAME] for increased vaccine access for South Philadelphia seniors and for members of our immigrant communities. We can’t let physical distance and language barriers keep people from this lifesaving vaccine.”* This tweet talks about vaccine access for the immigrant community but it doesn’t hold any stance towards any immigration policy.


A.1.3 Stance: Abortion

We aim to study the topic and the stance of tweets. Given a tweet accompanying with an image, you need to answer the stance of the tweet towards a given topic, as depicted in Figure 23. Please make sure that you have the basic knowledge about that social topic and you understand the key message that the tweet (i.e. both the text and the image) sends. Just skip the HIT if you are not sure.

The question is about the stance. You need to decide whether the tweet is relevant or not to the social topic abortion. If it is relevant, then you need to annotate the stance: supports/opposes to/doesn't hold any stance.

A tweet is considered as relevant if it talks about anything that discusses whether the abortion should be a legal option. If the arguments in the tweet text and image support that the abortion should be a legal option, then please choose "supports"; if arguments oppose to legal abortion, then choose "opposes to"; if arguments doesn't hold any stance for the topic then choose "doesn't hold any stance". Notice that a tweet is considered as irrelevant if it doesn't directly discuss whether the abortion should be a legal option or not, even though it may talk about related topics such as babies born alive after an abortion, birth control, etc.

Texas Abortion Clinics: We Should be Able to Dismember Unborn Babies While Their Hearts are Still Beating



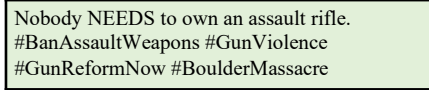
This tweet _____ the topic "Abortion".

- supports
- opposes to
- doesn't hold any stance to
- is not relevant to

Figure 23: Example of stance annotation on abortion.

A.2 Persuasiveness Level and Image Content

We aim to study the persuasiveness level of images in tweets as well as their content. Given a tweet text shown as Figure 24, you need to give a persuasiveness score of it. Then given a tweet accompanying an image shown as Figure 25, you need to give a persuasiveness score again.

A rectangular box with a black border containing text. The text is: "Nobody NEEDS to own an assault rifle. #BanAssaultWeapons #GunViolence #GunReformNow #BoulderMassacre".

Nobody NEEDS to own an assault rifle.
#BanAssaultWeapons #GunViolence
#GunReformNow #BoulderMassacre

Figure 24: Example of a text only tweet.

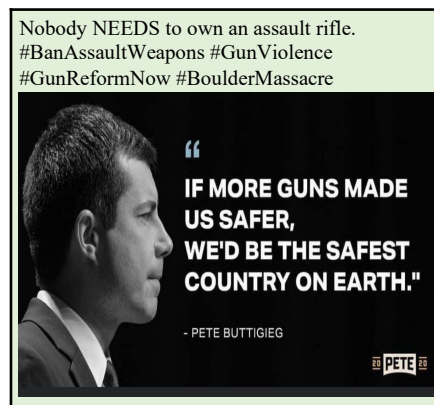


Figure 25: Example of a tweet accompanying an image.

Finally, you need to select the content type of the image. The content type of an image represents what type of the information the image mainly carries. Specifically, you need to pick one out of six types below for each image.

Statistics: the image provides evidence by **stating or quoting quantitative information**, such as a chart/data analysis, that is related to the tweet text.

An image could be considered statistics if: 1) It carries quantitative information (number/statistics/etc). 2) The key purpose of the image is to deliver this quantitative information, in the case there are multiple content types involved.


<p>Statistics: Compared to other developed countries the US suffers from higher gun fatalities than many other countries it has a more than 3 times the amount of deaths...</p>	<p>NOT Statistics: America has a #GunViolence problem the manufacturers make money hand over fist, funnel millions into the #GOP and we loose lives and loved ones...</p>																														
 <table border="1"> <caption>Homicides by firearm per 1 million people</caption> <thead> <tr> <th>Country</th> <th>Rate (per 1 million people)</th> </tr> </thead> <tbody> <tr><td>Australia</td><td>1.4</td></tr> <tr><td>New Zealand</td><td>1.6</td></tr> <tr><td>Germany</td><td>1.9</td></tr> <tr><td>Austria</td><td>2.2</td></tr> <tr><td>Denmark</td><td>2.7</td></tr> <tr><td>Netherlands</td><td>3.3</td></tr> <tr><td>Sweden</td><td>4.1</td></tr> <tr><td>Finland</td><td>4.5</td></tr> <tr><td>Ireland</td><td>4.8</td></tr> <tr><td>Canada</td><td>5.1</td></tr> <tr><td>Luxembourg</td><td>6.2</td></tr> <tr><td>Belgium</td><td>6.8</td></tr> <tr><td>Switzerland</td><td>7.7</td></tr> <tr><td>United States</td><td>29.7</td></tr> </tbody> </table> <p>SOURCE: UNODC, Small Arms Survey, via The Guardian.</p>	Country	Rate (per 1 million people)	Australia	1.4	New Zealand	1.6	Germany	1.9	Austria	2.2	Denmark	2.7	Netherlands	3.3	Sweden	4.1	Finland	4.5	Ireland	4.8	Canada	5.1	Luxembourg	6.2	Belgium	6.8	Switzerland	7.7	United States	29.7	 <p>DEVELOPING STORY AT LEAST 2 KILLED, 8 INJURED IN MULTIPLE SHOOTINGS IN VIRGINIA CNN LIVE NEWSROOM</p>
Country	Rate (per 1 million people)																														
Australia	1.4																														
New Zealand	1.6																														
Germany	1.9																														
Austria	2.2																														
Denmark	2.7																														
Netherlands	3.3																														
Sweden	4.1																														
Finland	4.5																														
Ireland	4.8																														
Canada	5.1																														
Luxembourg	6.2																														
Belgium	6.8																														
Switzerland	7.7																														
United States	29.7																														

Figure 26: Example of tweets with statistics image and a non-statistics image.

For the examples shown in Figure 26, in the statistics example, the image mainly shows a chart and delivers quantitative information (homicides by firearm per 1 million people). In contrast, in the NOT statistics example, though there are numbers in the image, the main information is a news title and the shooting scene, but not these numbers.

Testimony: the image **quotes statements or conclusions from an authority**, such as a piece of an article/claim from an official document, that is related to the tweet text.

The image can be considered as testimony if: 1) The content contains texts such as statements/conclusions/pieces of article. 2) These texts are original from other resources such as news/celebrities/official documents/etc. 3) The key purpose of the image is to quote the authorized statement, in the case there are multiple content types involved.

For the examples shown in Figure 27, in the Testimony tweet example, the image mainly cites a statement given by the transportation secretary. However, in the NOT Testimony tweet example, though it contains a piece of texts, these texts are not cited from an authority, therefore, it is not testimony.

Anecdote: the image provides information based on the **author's personal experience**, such as facts/personal stories, that are related to the tweet text.


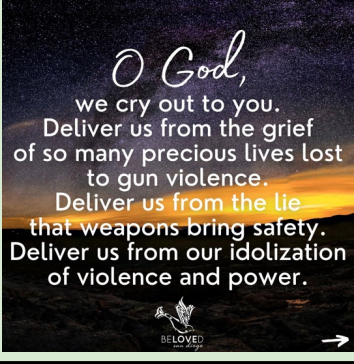
<p>Testimony: Nobody NEEDS to own an assault rifle. #BanAssaultWeapons #GunViolence #GunReformNow #BoulderMassacre</p> 	<p>NOT Testimony: Lord, make us instruments of your #Peace. Empower us to bring an end to #GunViolence, which has taken the lives of so many of your Beloved children</p> 
---	---

Figure 27: Example of tweets with testimony image and a non-testimony image.


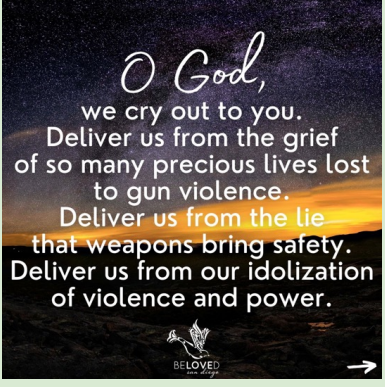
<p>Anecdote: Keep your guns but reform the #laws. During the founding fathers days #Guns were needed for protecting, hunting etc they didnt have to worry about over populated #malls, #terrorism etc...</p> 	<p>NOT Anecdote: Lord, make us instruments of your #Peace. Empower us to bring an end to #GunViolence, which has taken the lives of so many of your Beloved children.</p> 
---	---

Figure 28: Example of tweets with anecdote image and a non-anecdote image.

An image can be considered as an anecdote if: 1) It delivers a personal experience, Or 2) it shows a fact/experience that comes from personal view/known by the author. 3) The key purpose of the image is to deliver personal experience, in the case there are multiple content types involved.

For the examples shown in Figure 28, the anecdote image shows the personal view on the fact that guns have been developed since the period of the 2nd amendment, and therefore the laws for guns should be developed as well. However, in the NOT anecdote example, though it comes from a personal statement, it does not describe any fact/experience/stories.

Slogan: the image expresses a piece of **advertising phrase**.

An image can be considered as a slogan if: 1) It mainly delivers a piece of text as slogan; 2) The text is for advertising purposes as an advertising phrase/claim/statement. 3) The key purpose of the image is to deliver the piece of text, in the case there are multiple content types involved.

For the examples shown in Figure 29, the slogan image presents a phrase “Actually guns do kill people. Gun Reform Now”, therefore it is a slogan. However, For the example of NOT Slogan, though the image is for advertising, it does not contain a phrase for that, therefore it is not a slogan.

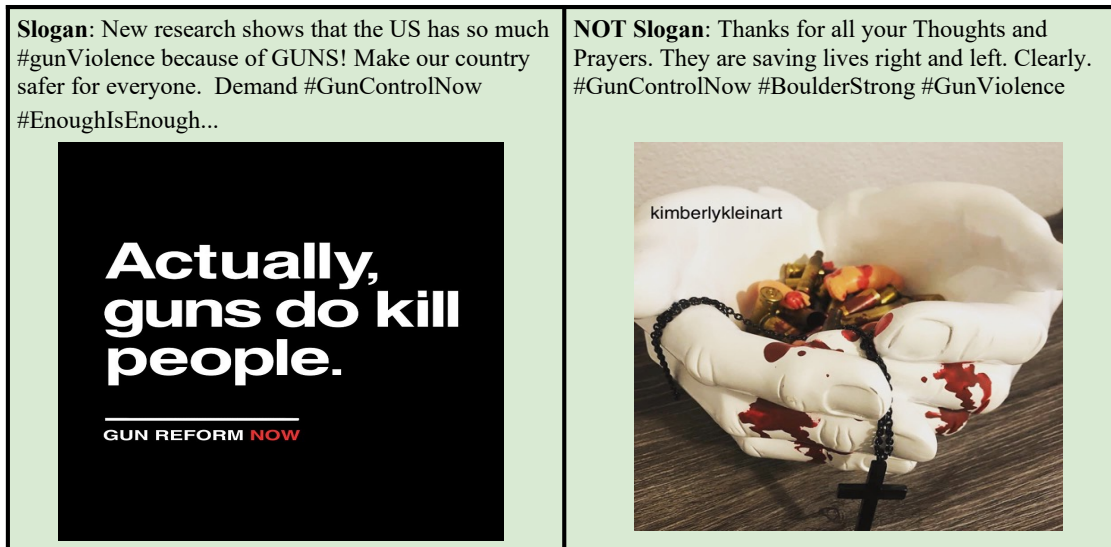


Figure 29: Example of tweets with slogan image and a non-slogan image.

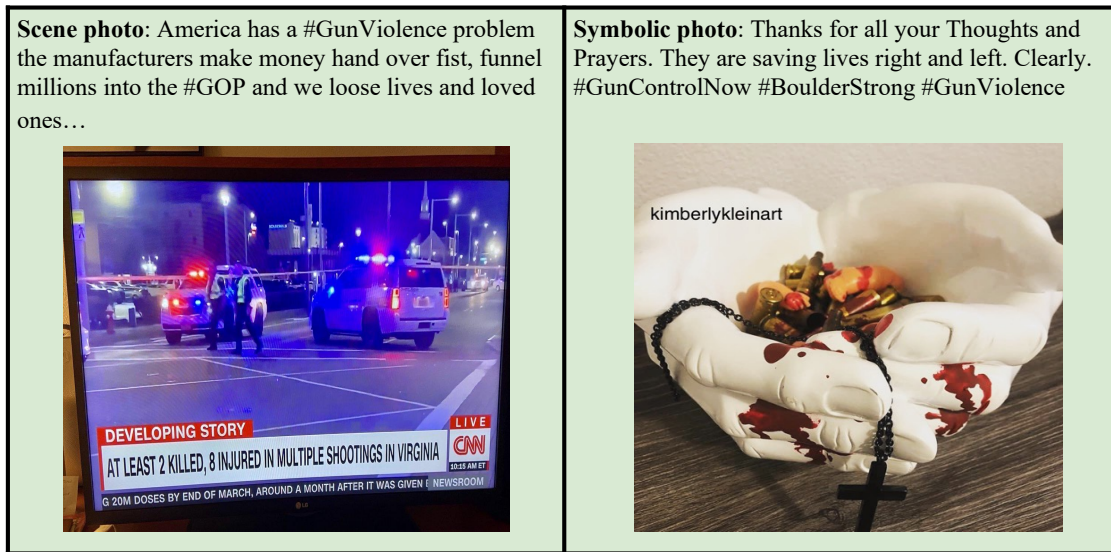


Figure 30: Example of tweets with scene photo image and a symbolic photo image.

Scene photo: the image shows a **literal scene/photograph** that is related to the tweet text.

An image can be considered as a scene photo if: 1) It shows a literal photograph/scene. 2) The image is directly related to the text. 3) The key purpose of the image is to deliver the image content but not the text within, in the case there are multiple content types involved.

Symbolic photo: the image shows a **symbol/art** that expresses the author's viewpoints in a **non-literal** way.

An image can be considered as a symbolic photo if: 1) It shows a symbol/art. 2) It expresses the viewpoint from the author in an implicit way. 3) The key purpose of the image is to deliver the image content but not the text within, in the case there are multiple content types involved.

For example, in Figure 30, the scene photo image shows a real photograph of a gun violence scene reported by CNN news. In the Symbolic photo, though relevant to the text, it shows a photo/image that is related to the text in a non-literal way (blood signifies gun-killing and the hand posture signifies praying), therefore it is not a scene photo but a symbolic photo.



Figure 31: Another example of tweets with scene photo image and a symbolic photo image.

The key difference between the Scene photo and Symbolic photo is **whether the photograph sends a message literally or symbolically**. For a scene photo, the image directly expresses/supports the author's view without any rhetoric; for a symbolic photo, the image may have several possible interpretations and the audience can understand its symbolic meaning after considering the tweet text. Consider the example shown in Figure 31: for the scene photo, it directly shows a protest scene and the author opposes to the abortion by considering it as a lie. In the symbolic photo, the author shows a photo of Notre Dame as a symbol of anti-abortion. The photo is not directly related to abortion, but audience can understand its symbolic meaning after reading the text.

In the case there are multiple content types involved: You need to first identify the key purpose of the image (i.e. what is the most important information in the image). Then please select the content type of the key purpose. Table 25 shows the summary of content types for each key purpose employed in the images.

Table 25: Summary of content types for each key purpose employed in the images.

Key Purpose		Content Type
Quantitative information in the image		Statistics
Textual information in the image	Statements or conclusions from an authority	Testimony
	Personal experiences/views	Anecdote
	Advertising phrases	Slogan
Graphical information in the image	Literal photograph	Scene Photo
	Non-literal/rhetorical photograph	Symbolic Photo

A.3 Persuasion Mode

We aim to study the **argumentative roles of images** in tweets. Given a tweet accompanying an image, we would ask you to choose the persuasion mode of the image. The persuasion mode of an image represents how the image convinces the audience. Specifically, we will ask you whether the image appeals to logic/emotion/credibility. Additionally, we will ask you why you make the choices.

Q1: Does the image make the tweet more persuasive by appealing to **logic and reasoning**?

The image appeals to logic and reasoning if it persuades audiences with reasoning from a fact/statistics/study case/scientific evidence. Specifically, if: 1) the image **contains information for logic and reasoning**; 2) the image **presents logic and reasoning**.

Also, we will ask you why you made the choice. i.e. Describing the logic/reasoning brought by the image. Such as following, by filling the blank in the textbox:

The logic/reasoning of the image is [the correlation between gun deaths and gun ownership by population].

For example shown in Figure 32, the left image provides a chart that shows the high gun deaths and the high gun ownership by the population of the US, which implies [a correlation between gun death and gun ownership which demonstrates that there will be less gun deaths

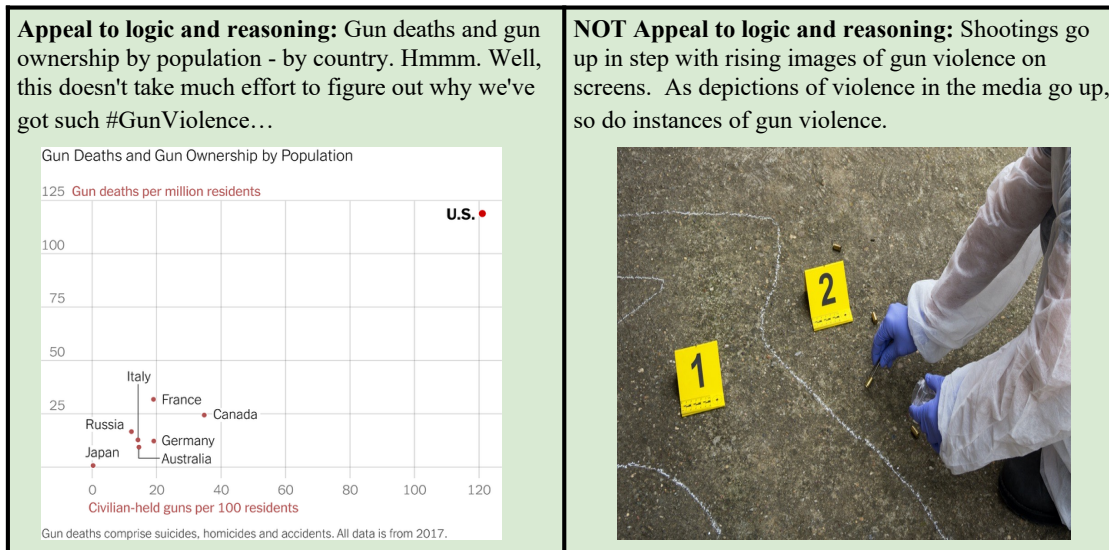


Figure 32: Example of tweets with logos image and non-logos image.

with gun control.]. On the contrary, the right image shows the scene of the shooting but does not provide any reasoning or logic.

Q2: Does image make the tweet more persuasive by appealing to **emotion**?

The image appeals to **emotion**, if it puts audiences in a certain frame of mind by stimulating them to identify/empathize/sympathize with the arguments.

Specifically, if : 1) the image **invokes the audience with strong emotion**, such as sadness, happiness, compassion, worry; 2) the image **makes the audience identify/empathize/sympathize** with the author/arguments.

Also, we will ask you why you made the choice. i.e. Describing the emotion(such as anger/amusement/sad/etc.) or impulsion(desire to do something) brought by the image. Such as following, by filling the blank within the [bracket]:

The image evokes my emotion/impulse of [anger].

For example shown in Figure 33, the left image shows the grieving "Uncle Sam" saying "no" with helplessness, which evokes the [desire for gun control]. The right image provides an item that can revoke [compassion and forgiveness].

Q3: Does image make the tweet more persuasive by **enhancing credibility and trust-**



Figure 33: Example of tweets with pathos images.



Figure 34: Example of tweets with ethos image and non-ethos image.

worthiness?

The image **enhances credibility and trustworthiness**, if it makes people trust something more via authorized/trusted expertise/title/reputation.

Specifically, if 1) The image **cites reliable sources** of the event/story/opinion/stance, that can make the contents trustworthy. Reliable sources include news, research reports, celebrated dictum, etc. Sources which are not proved/well-known by the audience (.e.g. an organization logo) are not considered as reliable. 2) the image **shows authorities** that can convince the audience to believe the arguments.

Also, we will ask you why you made the choice. i.e. Describing the resources of the citation that enhances the credibility. Such as following, by filling the blank within the [bracket]:

The credibility is enhanced by [a citation to political report]

For example shown in Figure 34, the left image takes a screenshot of the source of a report from [New York Times], which increases credibility. The NOT Ethos right image shows the views but are not quoted sentences that do not provide the credibility to enhance the argument.

Appendix B Annotation Instruction for SymbA

*Please describe the object which is in the red box.

*The description should be 1) in a short noun phrase, i.e. maximum 8 words (e.g. tooth under an umbrella); 2) capable to tell its symbolic meaning that is already given (e.g. blood signifies danger; lemon signifies refreshing; tooth under an umbrella signifies protection and health).

*Instruction for corner cases:

1) If there are multiple objects in the red box, please first identify several objects which relate to the given symbolic meaning, then describe them and their relationship in a short phrase, e.g. tooth under an umbrella.

2) If some attributes of the target object is essential for telling its symbolic meaning, please describe the attribute (e.g. color, shape, status, action) with the class name together, e.g. bleeding arm

*In summary, the goal is to infer the given symbolic meaning from your written description. If you meet some cases which are not covered by the instruction, please write a description which helps most for inferring the given symbolic meaning.

*Some examples of expected annotations are shown on the first page of this form: [\[link\]](#)

Bibliography

- [1] Arjun R Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T Freeman, et al. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23201–23211, 2023.
- [2] Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, 2016.
- [3] Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. Cite: A corpus of image-text discourse relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 570–575, 2019.
- [4] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770, 2020.
- [5] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18, 2015.
- [6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.
- [7] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016.
- [8] George M Belknap. Communication and persuasion: Psychological studies of opinion change. by carl i. hovland, irving l. janis and harold h. kelley. *American Political Science Review*, 48(2):600–600, 1954.

- [9] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3286–3295, 2019.
- [10] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9592–9600, 2019.
- [11] Gedas Bertasius and Lorenzo Torresani. Cobe: Contextualized object embeddings from narrated instructional video. *Advances in Neural Information Processing Systems*, 33:15133–15145, 2020.
- [12] Douglas Brown. The penguin dictionary of symbols. *Reference Reviews*, 1997.
- [13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [14] Kenneth Burke. Rhetoric of hitler’s battle, the. *The Southern Review*, 5:1, 1939.
- [15] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.
- [16] Winston Carlile, Nishant Gurrupadi, Zixuan Ke, and Vincent Ng. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, 2018.
- [17] Shelly Chaiken. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology*, 39(5):752, 1980.
- [18] Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathleen Mckeown, and Alyssa Hwang. Ampersand: Argument mining for persuasive online discussions. 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 2933–2943, 2019.

- [19] Niladri Chatterjee and Saumya Agrawal. Word alignment in english-hindi parallel corpus using recency-vector approach: some studies. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 649–656, 2006.
- [20] Jiaao Chen and Diyi Yang. Weakly-supervised hierarchical models for predicting persuasive strategies in good-faith textual requests. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12648–12656, 2021.
- [21] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [22] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020.
- [23] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7):853–862, 2012.
- [24] J Cohen. Statistical power analysis for the behavioral sciences lawrence earlbaum associates. 20th–, 1988.
- [25] Edward Meredith Cope and John Edwin Sandys. *The rhetoric of Aristotle*, volume 2. University Press, 1877.
- [26] Joe Davison, Joshua Feldman, and Alexander Rush. Commonsense knowledge mining from pretrained models. 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 1173–1178, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings*

- of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, 2019.
- [29] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1422–1430, 2015.
- [30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.
- [31] Rory Duthie, Katarzyna Budzynska, and Chris Reed. Mining ethos in political debate. In *Computational Models of Argument: Proceedings from the Sixth International Conference on Computational Models of Argument (COMMA)*, pages 299–310. IOS Press, 2016.
- [32] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [33] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.
- [34] Nancy Green, Kevin D Ashley, Diane Litman, Chris Reed, and Vern Walker. Proceedings of the first workshop on argumentation mining. In *Proceedings of the First Workshop on Argumentation Mining*, 2014.
- [35] Meiqi Guo, Rebecca Hwa, and Adriana Kovashka. Detecting persuasive atypicality by modeling contextual compatibility. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 972–982, 2021.
- [36] Meiqi Guo, Rebecca Hwa, and Adriana Kovashka. Decoding symbolism in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3311–3324, 2023.

- [37] Meiqi Guo, Rebecca Hwa, Yu-Ru Lin, and Wen-Ting Chung. Inflating topic relevance with ideology: A case study of political ideology bias in social topic detection models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4873–4885, 2020.
- [38] Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179, 2017.
- [39] Edward L Hancock. *Techniques for Understanding Literature: A Handbook for Readers and Writers*. Wadsworth Publishing Company, 1972.
- [40] Lisa Hartling, Michele Hamm, Andrea Milne, Ben Vandermeer, P Lina Santaguida, Mohammed Ansari, Alexander Tsertsvadze, Susanne Hempel, Paul Shekelle, and Donna M Dryden. Validity and inter-rater reliability testing of quality assessment instruments. 2012.
- [41] Adi Haviv, Jonathan Berant, and Amir Globerson. BERTese: Learning to speak to BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3618–3623, Online, April 2021. Association for Computational Linguistics.
- [42] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.
- [43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [44] Benjamin Heinzerling and Kentaro Inui. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online, April 2021. Association for Computational Linguistics.
- [45] Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathleen McKeown. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, 2017.

- [46] Colin Higgins and Robyn Walker. Ethos, logos, pathos: Strategies of persuasion in social/environmental reports. In *Accounting forum*, volume 36, pages 194–208. Elsevier, 2012.
- [47] Nigel Hollis. Why good advertising works (even when you think it doesn’t). *The Atlantic*, 31, 2011.
- [48] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer: Generating music with long-term structure. In *International Conference on Learning Representations*, 2018.
- [49] Xinyue Huang and Adriana Kovashka. Inferring visual persuasion via body language, setting, and deep features. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 778–784, 2016.
- [50] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1705–1715, 2017.
- [51] Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. Talking to the crowd: What do people react to in online discussions? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2026–2031, 2015.
- [52] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.
- [53] Ernest Jones. The theory of symbolism. *British Journal of Psychology*, 9(2):181, 1918.
- [54] Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu. Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–223, 2014.
- [55] Yash Kant, Dhruv Batra, Peter Anderson, Alex Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. Spatially aware multimodal transformers for textvqa. In *European Conference on Computer Vision*, pages 715–732. Springer, 2020.

- [56] Walter Kintsch. Metaphor comprehension: A computational theory. *Psychonomic bulletin & review*, 7(2):257–266, 2000.
- [57] Klaus Krippendorff. Computing krippendorff’s alpha-reliability. 2011.
- [58] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.
- [59] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. Integrating text and image: Determining multimodal document intent in instagram posts. 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 4622–4632, 2019.
- [60] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.
- [61] Ron Langacker. Cognitive linguistics symposium. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society: July 12-15, 1996, University of California, San Diego*, volume 18, page 15. Psychology Press, 1996.
- [62] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European Conference on Computer Vision*, pages 577–593. Springer, 2016.
- [63] John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818, 2020.
- [64] Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE international symposium on circuits and systems*, pages 253–256. IEEE, 2010.
- [65] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020.

- [66] Hongsong Li, Kenny Q. Zhu, and Haixun Wang. Data-driven metaphor recognition and explanation. *Transactions of the Association for Computational Linguistics*, 1:379–390, 2013.
- [67] Kun-Lun Li, Hou-Kuan Huang, Sheng-Feng Tian, and Wei Xu. Improving one-class svm for anomaly detection. In *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 03EX693)*, volume 5, pages 3077–3081. IEEE, 2003.
- [68] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
- [69] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, August 2021. Association for Computational Linguistics.
- [70] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.
- [71] Yiyi Li and Ying Xie. Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of Marketing Research*, 57(1):1–19, 2020.
- [72] Karin Liebhart and Petra Bernhardt. Political storytelling on instagram: Key aspects of alexander van der bellen’s successful 2016 presidential election campaign. *Media and Communication*, 5(4):15–25, 2017.
- [73] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [74] Zhexiong Liu, Meiqi Guo, Yue Dai, and Diane Litman. ImageArg: A multi-modal tweet dataset for image persuasiveness mining. In *Proceedings of the 9th Workshop on Argument Mining*, pages 1–18, Online and in Gyeongju, Republic of Korea, October 2022. International Conference on Computational Linguistics.

- [75] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32:13–23, 2019.
- [76] Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, 2017.
- [77] Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, and Kalina Bontcheva. The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management*, 56(6):102055, 2019.
- [78] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.
- [79] Tanushree Mitra and Eric Gilbert. The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 49–61, 2014.
- [80] Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23, 2017.
- [81] Pedro Morgado, Yi Li, and Nuno Vasconcelos. Learning representations from audio-visual spatial alignment. In *NeurIPS*, 2020.
- [82] Caroline Lego Munoz and Terri L Towner. The image is the message: Instagram marketing and the 2016 presidential primary season. *Journal of political marketing*, 16(3-4):290–318, 2017.
- [83] Arthur Neidlein, Philip Wiesenbach, and Katja Markert. An analysis of language models for metaphor recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3722–3736, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.
- [84] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

- [85] Daniel J O’keefe. *Persuasion: Theory and research*. Sage Publications, 2015.
- [86] OpenAI. Gpt-4 technical report, 2023.
- [87] Sunghyun Park, Han Suk Shim, Moitreyia Chatterjee, Kenji Sagae, and Louis-Philippe Morency. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 50–57, 2014.
- [88] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018.
- [89] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.
- [90] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [91] Isaac Persing and Vincent Ng. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, 2015.
- [92] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [93] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? 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 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics.

- [94] Richard E. Petty and John T. Cacioppo. *The Elaboration Likelihood Model of Persuasion*. Springer New York, New York, NY, 1986.
- [95] Richard E Petty and John T Cacioppo. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer Science & Business Media, 2012.
- [96] Ivan S Pimenov, Natalia V Salomatina, and Mariya K Timofeeva. The quantitative evaluation of the pathos to ethos ratio in scientific texts. In *2022 IEEE 23rd International Conference of Young Professionals in Electron Devices and Materials (EDM)*, pages 312–317. IEEE, 2022.
- [97] Reid Pryzant, Youngjoo Chung, and Dan Jurafsky. Predicting sales from the language of product descriptions. In *eCOM@ SIGIR*, 2017.
- [98] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.
- [99] Guanghai Qin and Jason Eisner. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online, June 2021. Association for Computational Linguistics.
- [100] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [101] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [102] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.
- [103] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

- [104] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, 32, 2019.
- [105] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28:91–99, 2015.
- [106] Zachary Rosen. Computationally constructed concepts: A machine learning approach to metaphor interpretation using usage-based construction grammatical cues. In *Proceedings of the Workshop on Figurative Language Processing*, pages 102–109, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [107] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
- [108] Babak Saleh, Ahmed Elgammal, Jacob Feldman, and Ali Farhadi. Toward a taxonomy and computational models of abnormalities in images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.
- [109] Andrey Savchenko, Anton Alekseev, Sejeong Kwon, Elena Tutubalina, Evgeny Myasnikov, and Sergey Nikolenko. Ad lingua: Text classification improves symbolism prediction in image advertisements. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1886–1892, 2020.
- [110] Robin Schaefer and Manfred Stede. Argument mining on twitter: A survey. *it-Information Technology*, 63(1):45–58, 2021.
- [111] Martijn Schoonvelde, Anna Brosius, Gijs Schumacher, and Bert N Bakker. Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches. *PloS one*, 14(2):e0208450, 2019.
- [112] Linda M Scott. Images in advertising: The need for a theory of visual rhetoric. *Journal of consumer research*, 21(2):252–273, 1994.
- [113] Omar Shaikh, Jiaao Chen, Jon Saad-Falcon, Polo Chau, and Diyi Yang. Examining the ordering of rhetorical strategies in persuasive requests. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1299–1306, 2020.

- [114] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, 2018.
- [115] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, 2020.
- [116] Ekaterina Shutova. Automatic metaphor interpretation as a paraphrasing task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037, Los Angeles, California, June 2010. Association for Computational Linguistics.
- [117] Ekaterina Shutova, Douwe Kiela, and Jean Maillard. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 160–170, 2016.
- [118] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [119] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017.
- [120] Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In *ArgNLP*, pages 21–25, 2014.
- [121] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, 2018.
- [122] Gerard J Steen. *Visual metaphor: Structure and process*, volume 18. John Benjamins Publishing Company, 2018.

- [123] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019.
- [124] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.
- [125] Arthur Symons. *The symbolist movement in literature*. Carcanet, 2014.
- [126] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [127] 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*, pages 5103–5114, 2019.
- [128] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.
- [129] Tony Veale and Yanfen Hao. A fluid knowledge representation for understanding and generating creative metaphors. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 945–952, Manchester, UK, August 2008. Coling 2008 Organizing Committee.
- [130] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.
- [131] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *European Conference on Computer Vision*, pages 391–408. Springer, 2018.

- [132] Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al Khatib, Maria Skeppstedt, and Benno Stein. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753–3765, 2018.
- [133] Alex Wang and Kyunghyun Cho. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [134] Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.
- [135] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1386–1393, 2014.
- [136] Peng Wang, Lingqiao Liu, Chunhua Shen, Zi Huang, Anton van den Hengel, and Heng Tao Shen. What’s wrong with that object? identifying images of unusual objects by modelling the detection score distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1573–1581, 2016.
- [137] Xuwei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, 2019.
- [138] Zhongyu Wei, Yang Liu, and Yi Li. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200, 2016.
- [139] Judith Williamson. *Decoding advertisements: ideology and meaning in advertising*. Marion Boyers, 1978.
- [140] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

- [141] Muheng Yan, Yu-Ru Lin, Rebecca Hwa, Ali Mert Ertugrul, Meiqi Guo, and Wen-Ting Chung. Mimicprop: Learning to incorporate lexicon knowledge into distributed word representation for social media analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 738–749, 2020.
- [142] Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. Let’s make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3620–3630, 2019.
- [143] Keren Ye, Narges Honarvar Nazari, James Hahn, Zaeem Hussain, Mingda Zhang, and Adriana Kovashka. Interpreting the rhetoric of visual advertisements. *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- [144] Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. Multimet: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3214–3225, 2021.
- [145] Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *Proceedings of the British Machine Vision Conference (BMVC)*, September 2018.
- [146] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.
- [147] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016.
- [148] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.
- [149] Xizhou Zhu, Dazhi Cheng, Zheng Zhang, Stephen Lin, and Jifeng Dai. An empirical study of spatial attention mechanisms in deep networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6688–6697, 2019.

- [150] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.
- [151] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.