Benchmarking VLMs' Reasoning About Persuasive Atypical Images

Sina Malakouti[1,*] Aysan Aghazadeh[1,*] Ashmit Khandelwal[2] Adriana Kovashka[1]
[1] University of Pittsburgh [2]BITS Pilani
{sem238, aya34}@pitt.edu ashmitk0507@gmail.com kovashka@cs.pitt.edu

Abstract

Vision language models (VLMs) have shown strong zero-shot generalization across various tasks, especially when integrated with large language models (LLMs). However, their ability to comprehend rhetorical and persuasive visual media, such as advertisements, remains understudied. Ads often employ ***atypical imagery****, using surprising object juxtapositions to convey shared properties. For example, Fig. 1 (e) shows a beer with a feather-like texture. This requires* ***advanced reasoning*** *to deduce that this atypical representation signifies the beer's lightness.*

We introduce three novel tasks, Multi-label Atypicality Classification, Atypicality Statement Retrieval, and Atypical Object Recognition, to benchmark VLMs' understanding of atypicality in persuasive images. We evaluate how well VLMs use atypicality to infer an ad's message and test their reasoning abilities by employing semantically challenging negatives. Finally, we pioneer atypicality-aware verbalization by extracting comprehensive image descriptions sensitive to atypical elements.

Our findings reveal that: (1) VLMs lack advanced reasoning capabilities compared to LLMs; (2) simple, effective strategies can extract atypicality-aware information, leading to comprehensive image verbalization; (3) atypicality aids persuasive advertisement understanding. Code and data are available at github.com/sinamalakouti/PersuasiveAdVLMBenchmark

1. Introduction

In visual media, particularly advertisements, creators employ *creative* visual rhetoric to capture attention and convey memorable, powerful messages. They intentionally deviate from realism, depicting objects in unique and atypical ways [29, 34]. Creative ads that are "out of the ordinary" or "connect objects that are usually unrelated" can generate twice as much revenue as non-creative ads [34].

Atypical imagery in ads often involves transforming objects metaphorically [43, 46]. These creative transformations are not random; they are carefully chosen to convey specific ideas [43]. For example, Fig. 1(a) depicts a text box as tape to suggest silencing, while in (d), potato chips are shown as flames to metaphorically represent spiciness, borrowing properties from fire (hotness symbolizing flavor). Understanding these atypical images requires more than just recognizing objects. It requires advanced reasoning skills, including knowledge of cultural contexts and social norms, posing a significant challenge for AI systems.

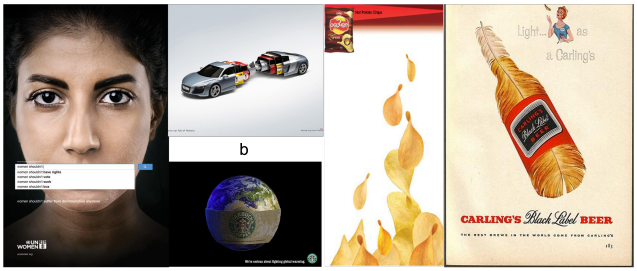

Figure 1. **Atypicality categories.** We study four types of atypicality from [46]: Texture Replacement 1, Texture Replacement 2, Object Inside Objects, Object Replacement (defined in Sec. 3.1).

Modern pretrained vision-language models (VLMs), such as LLaVA [26, 27], demonstrate strong visual understanding across various tasks such as recognition [25], and capabilities like zero-shot generalizability. However, there is a lack of in-depth study on VLMs' ability to understand complex persuasive images, such as advertisements.

We address this gap by introducing three novel tasks over PittAds [18] to evaluate VLMs' understanding of atypicality: (1) multi-label atypicality classification, where the model predicts the type of atypicalities in the image; (2) atypicality statement retrieval, where the model retrieves correct atypicality statements describing the atypicality relation among objects; (3) atypical object recognition, where the model generates objects to complete an atypicality statement based on a given relation. These tasks are essential as prior works' binary classification oversimplifies atypicality's nuanced nature. Our evaluation shows that although

*These authors contributed equally to this work.