# Do Vision–Language Models Understand Visual Persuasiveness? A Diagnosis via Visual Persuasive Factors

**Anonymous ACL submission**

## Abstract

Visual persuasion uses images to shape cognition, emotion, and behavior, with its effects depending on both visual attributes and semantic context. Despite recent progress, it remains unclear whether Vision–Language Models (VLMs) understand visual persuasiveness. This motivates us to ask: can VLMs assess whether an image persuasively supports an intended message, which visual factors shape this judgment, and do they align with human judgments? Through empirical analyses on image-message pairs where human raters consistently agree on the persuasiveness judgment, we show that VLMs exhibit a recall-oriented bias: they over-predict images as persuasive while achieving high recall. We introduce **Visual Persuasive Factors (VPFs)**, a taxonomy informed by cognitive psychology for quantifying visual cues that shape persuasive judgments. Our factor-level analysis reveals that VPFs distinguish human persuasiveness judgments, whereas VLMs only partially reproduce these patterns—often generating false positives by treating persuasion-relevant cues as sufficient evidence. Building on this insight, we evaluate VPF-guided interventions and find that properly framed VPF knowledge can improve performance, but merely specifying visual cues or adding step-by-step reasoning is insufficient. By analyzing model rationales at the level of functional reasoning steps, we further identify a central bottleneck in connecting object identification to semantic message alignment.[1]

## 1 Introduction

Visual persuasion refers to the use of images and visual elements to influence cognition, emotion, and behavior, and it plays a central role in advertising, political communication, and social media (Chandler and Munday, 2011; Joo et al., 2014; Kim et al., 2025). Visuals are not just decorative—they reliably shape attitudes and behavior, as classic persuasion theory (Petty and Cacioppo, 1986) and public-health evidence on pictorial warnings have shown (Noar et al., 2016; Brewer et al., 2016). Visual elements are therefore central to persuasion, but their effects unfold through their interaction with semantic context, persuasion strategies, and individual psychological dispositions (Kumar et al., 2023; Kim et al., 2025). However, existing work provides limited insight into how well VLMs actually understand visual persuasiveness in this broader context—why certain images are persuasive, which visual factors contribute to their effects, and how these factors differ from those underlying human judgments remain unclear.

In this work, we address this gap by introducing a framework for examining the role of persuasion-relevant visual factors and comparing their associations with VLM and human judgments. To isolate the effect of visual factors, we control for the inherent subjectivity of persuasiveness judgments (Matz et al., 2024). Specifically, we use the Personalized Visual Persuasion (PVP) dataset (Kim et al., 2025), which provides image-message pairs with multiple human persuasiveness ratings, and retain 562 clearly separable persuasive (High) and non-persuasive (Low) binary cases with almost-perfect agreement (Landis and Koch, 1977). We then evaluate whether VLMs can assess how persuasively an image supports a given target message, as shown in Fig. 1 (Left). Surprisingly, we find that even in these seemingly straightforward scenarios, VLMs struggle to exhibit genuine understanding of visual persuasiveness; they tend to over-predict persuasiveness, revealing a pronounced *recall-oriented bias*. Together with insights from the Elaboration Likelihood Model (Petty and Cacioppo, 1986), we hypothesize that VLM limitations stem from how models identify, weight, and contextualize visual and semantic cues that contribute to persuasiveness.

To make this hypothesis operational, we propose **Visual Persuasive Factors (VPFs)**, a novel

[1]The dataset and code will be publicly released.

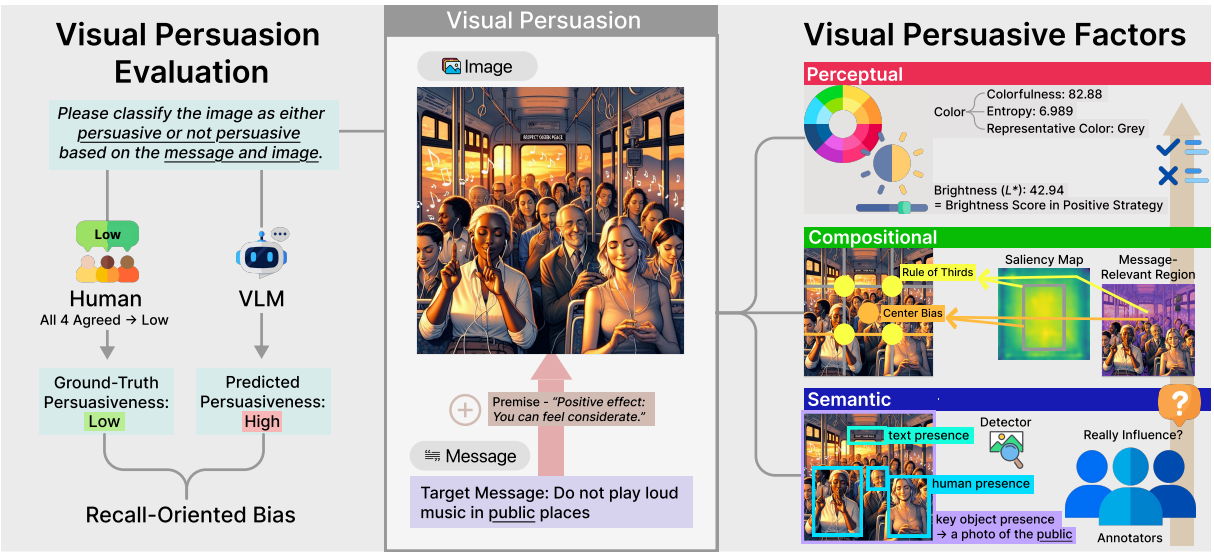



Figure 1: (Left) VLM persuasiveness judgments are evaluated against high-consensus human labels, revealing a recall-oriented bias. (Right) Visual Persuasive Factors span perceptual, compositional, and semantic cues, enabling factor-level analysis of visual persuasion. This setting and VPF construction are detailed in §3.1–§3.4.

structured taxonomy informed by cognitive psychology (Adams and Osgood, 1973; Janiszewski, 1990; Elder and Krishna, 2012) for measuring cues that shape persuasiveness judgments in image-message pairs, as illustrated in Fig. 1 (Right). VPFs span three levels—perceptual, compositional, and semantic—covering color and brightness; saliency and message-relevant region placement with respect to the center and rule of thirds; and the presence of key object, human, and text. A validation survey with 141 annotators, who are distinct from the original PVP raters, provides evidence for the validity of these factors. Using VPFs, we compare human ground-truth persuasion labels with VLM predictions, revealing three key insights:

- VPFs differentiate high- and low-persuasiveness cases based on interpretable visual and semantic dimensions.
- VLMs only partially reproduce human-aligned VPF patterns, capturing some persuasive signals while showing inconsistent weighting across models and distorted associations with factors such as human and text presence.
- VLM false positives arise from overgeneralizing persuasion-relevant cues: among ground-truth Low items, models often predict High when images contain plausible persuasive cues—*e.g.*, key object presence—mistaking them for sufficient evidence of persuasiveness.

Leveraging these insights, we further examine whether VPF-guided interventions can reduce the human–VLM gap in persuasive judgment. Focusing on key object presence as a semantic cue that reliably separates human judgments yet also contributes to VLM false positives, we test whether knowledge injection and reasoning scaffolds can help VLMs use persuasion-relevant cues more appropriately. We find that VLMs can improve when such cues are explicitly framed as auxiliary cues to be interpreted in context rather than as fixed evidence for persuasiveness. However, naive cue provision or explicit reasoning (Wei et al., 2022) is not sufficient for robust improvement: direct grounding often degrades performance, and reasoning with explicit cues can even amplify the recall-oriented bias, suggesting a failure to reliably connect visual cues to the final persuasive judgment. To better understand this limitation, we extract and decompose model-generated rationales into three functional steps—(i) object identification, (ii) semantic message support, and (iii) final judgment connection—and ablate each step to examine how it affects model predictions. This step-wise ablation shows that the difficulty lies in linking object-level and semantic evidence to communicative intent.

Our contributions are threefold: (1) we propose Visual Persuasive Factors (VPFs), a taxonomy informed by cognitive psychology, as a structured framework for visual persuasive understanding; (2) we use VPFs as a cognitive factor lens to compare human-labeled and VLM-predicted persuasiveness, showing that VPFs distinguish human judgments while VLMs only partially reproduce these human-aligned patterns and overgeneralize persuasion-relevant cues; and (3) we conduct VPF-

guided intervention and reasoning analyses, revealing that semantic grounding alone does not close the gap and that the key bottleneck lies in connecting visual evidence to communicative intent.

## 2 Related Work

### 2.1 AI-based Persuasion

Research on computational persuasion began in text-only dialogue settings, with corpora such as Persuasion for Good (Wang et al., 2019) enabling strategy-based persuasive response generation; later work incorporated affect, user traits, and LLM-generated messages to study empathy, personalization, and human judgments of persuasiveness (Samad et al., 2022; Matz et al., 2024; Bai et al., 2025a). In visual persuasion, early work used syntactic visual cues to infer communicative intents in persuasive images (Joo et al., 2014), while recent datasets such as PVP pair 28,454 persuasive images with 596 messages annotated for nine persuasion strategies (Kim et al., 2025). While these studies establish visual persuasion as a computational task, they leave open how VLM judgments relate to the visual factors that humans use to assess persuasiveness. Recent evaluation protocols measure persuasiveness more directly, including human-in-the-loop and controlled prompting setups for assessing model persuasion and safety implications (Durmus et al., 2024; Singh et al., 2024). Our study extends this paradigm to the visual domain by using factor-level analysis to compare model and human decisions.

### 2.2 Cognitive Foundations for Visual Persuasive Factors

Our taxonomy draws on cognitive psychology and communication theory. The Elaboration Likelihood Model (Petty and Cacioppo, 1986) posits two paths to persuasion: a central route operating via message-relevant elaboration, and a peripheral route relying on immediate or heuristic cues.

Numerous behavioral studies demonstrate that color is a fundamental component of visual perception and persuasion (Adams and Osgood, 1973; Mapelli and Behrmann, 1997; Gegenfurtner and Rieger, 2000; Labrecque and Milne, 2011), and is especially influential in advertisements and campaigns (Labrecque et al., 2013; Jr. et al., 2018; Armstrong et al., 2021). Beyond color, composition and visual hierarchy guide attention, as shown by center bias in natural scenes (Tatler, 2007; Zhang et al., 2008) and the rule-of-thirds in photographic composition (Krages, 2005; Mai et al., 2011). Advertising and eye-tracking studies further show that visual layout shapes gaze allocation and brand responses, highlighting composition as an attention-guiding factor in persuasive media (Janiszewski, 1990; Maughan et al., 2007). At the semantic level, persuasiveness depends on whether image objects align with the intended argument and form a coherent narrative, requiring models to select relevant objects and connect them to an argument structure (Chung et al., 2024). Such core objects may include products (Elder and Krishna, 2012), text (Zhang et al., 2018; Seo, 2020), or people (Selove, 2014; Meng et al., 2026).

### 2.3 High-Level Vision–Language Reasoning

Recent foundation models (OpenAI, 2025; Comanici et al., 2025; Guo et al., 2025) have advanced rapidly on complex tasks, while chain-of-thought prompting improves reasoning in arithmetic, commonsense, and symbolic settings (Wei et al., 2022; Wang et al., 2023; Yasunaga et al., 2024). VLMs extend this paradigm by coupling LLMs with visual encoders for multimodal inputs (Liu et al., 2023, 2024; Bai et al., 2025b; Team et al., 2025). Recent VLM research has moved beyond object recognition toward abstract and symbolic reasoning, including visual metaphor understanding, where models must bridge perception and semantics (Akula et al., 2023; Tong et al., 2024).

## 3 From VLM Persuasive Understanding to Visual Persuasive Factors

The first goal of our study is to examine whether VLMs possess a meaningful notion of visual persuasive understanding. In the following sections, we establish the evaluation setting using a high-agreement subset (§3.1), introduce our novel taxonomy Visual Persuasive Factors (§3.2), describe how factors are extracted and measured (§3.3), and validate them through an independent survey (§3.4).

### 3.1 Evaluating VLMs' Understanding of Visual Persuasiveness

To assess whether VLMs understand visual persuasion, we derive a high-consensus subset from the PVP dataset (Kim et al., 2025). We bin the original 0–10 persuasiveness ratings into Low (0–2), Mid (3–7), and High (8–10) bands, and retain 562 image-message pairs with four-annotator agreement in the Low or High band, *i.e.*, *almost-perfect*

agreement under Landis and Koch (1977). We then frame persuasion as a binary judgment task (image + message $\rightarrow$ {High, Low} treated as ground truth; details in Appendix A). Hence, this subset represents a distribution in which persuasive and non-persuasive cases are relatively clear to viewers. Yet VLMs (OpenAI, 2025; Team et al., 2025; Bai et al., 2025b) consistently show poor performance and a recall-oriented bias: they predict High too often, achieving near-perfect recall with inflated false positives, as shown in Tab. 2 baseline.[2]

### 3.2 Defining Visual Persuasive Factors

Building on prior findings in §2.2, we define Visual Persuasive Factors as a three-level decomposition of persuasive content in image-message pairs.

- **Perceptual** factors capture immediate visual impressions, including colorfulness, color entropy, representative color congruence, and brightness.
- **Compositional** factors describe the spatial organization of persuasive cues, including how visual saliency and message-relevant regions are distributed with respect to center bias and rule-of-thirds structure.
- **Semantic** factors capture various objects, including whether a key object, human, or text is present and whether they provide evidence that modulates the intended persuasive appeal.

Our VPFs are not intended to be exhaustive; they organize visually measurable cues into three diagnostic levels for VLM evaluation.

### 3.3 VPF Feature Extraction

**Perceptual Level.** We characterize perceptual factors in two complementary ways. First, we extract image-intrinsic properties that can be directly quantified, including *colorfulness, color entropy, and brightness*. We use the Hasler and Suesstrunk (2003) colorfulness metric to measure color intensity, compute color entropy in CIELAB space to capture color diversity, and use the $L^*$ channel to quantify brightness (McLAREN, 1976).

Second, we measure whether perceptual-level properties align with semantic framing. The intuition is that persuasiveness increases when cues implied by message intent are reflected in the image's visual attributes (*e.g.*, green color for environmental messages). For color, we first define a discrete eight-color affective palette inspired by Adams and Osgood (1973) and then extract a representative image color via K-means clustering and map it to one palette category. Separately, given the message and persuasion premise, GPT-5 infers the top-3 expected colors from the same palette.[3] Representative Color Congruence is the hit@3 overlap between the image's representative color and the inferred top-3 expected colors.

Similarly, for brightness, we score whether image brightness matches the implied persuasion strategy (*i.e.*, premise) by aligning $L^*$ values with strategy polarity (positive, negative, or neutral), following psychological patterns such as brighter images for positive strategies and darker images for negative ones (Adams and Osgood, 1973). In summary, we report normalized values for three color metrics (colorfulness, entropy, and representative color congruence), alongside raw brightness ($L^*$) and its strategy-specific alignment scores for positive- and negative-strategy subsets.[4] Full metric details and examples are in the Appendix B.1.

**Compositional Level.** To measure compositional factors, we focus on two canonical layout principles: the *Center Bias* (Zhang et al., 2008; Tatler, 2007) and the *Rule of Thirds* (RoT) (Krages, 2005). We evaluate them using two representations: log-probability saliency maps from DeepGaze IIE (Linardos et al., 2021), which capture visually attended regions, and Message-Relevant Region (MRR), defined as Grounded-SAM-2 (Ren et al., 2024) segmentation masks obtained from open-vocabulary queries derived from the message. For both representations, we compute Center Bias Index (CBI) and Thirds Hotspot Coverage (T3), which measure the proportion of total saliency mass or mask area contained in predefined circular regions around the image center and the four rule-of-thirds intersections, respectively. Higher values indicate stronger spatial concentration in those regions. Full metric details and examples are in Appendix B.2.

For perceptual and compositional factors, we compare each feature $f$ between persuasive and non-persuasive groups using

$$\Delta_f^{(Z)} = \mathbb{E}[f \mid Z = \text{High}] - \mathbb{E}[f \mid Z = \text{Low}], \quad (1)$$

[2] Although we evaluate LLaVA variants (Liu et al., 2023, 2024), we exclude them from the main analysis due to significant instruction-following failures and poor output stability.

[3] A premise is a short textual cue specifying the persuasion strategy for each image variation in PVP's one-message-to-many-images structure. An example is in Fig. 1.

[4] Since there were too few neutral cases, we only report the metrics for positive and negative cases.

where $Z \in \{Y, \hat{Y}\}$ denotes the ground-truth or the predicted label, respectively.

**Semantic Level.** We extract three factors from image content: *Key Object Presence*, *Human Presence*, and *Text Presence*. We operationalize all three factors in a simple binary form, focusing on whether each cue is present in the image. Key objects are extracted by identifying candidate nouns from the message and verifying whether their corresponding visual instances are depicted in the image; humans are detected using standard object detection; and text is manually verified for readable elements such as slogans or call-to-action phrases. In our curated dataset, the presence rates of these semantic factors are 56.58% for key objects, 71.00% for humans, and 22.06% for text. To evaluate their association, we fit logistic regressions with each semantic factor as a binary presence indicator and either ground-truth labels or model predictions as the outcome, reporting odds ratios (ORs). Full details of the pipeline are in Appendix B.3.

### 3.4 Validation Survey

We conduct an independent validation survey to assess the proposed VPFs by directly measuring their perceived influence on visual persuasive understanding. Annotators were recruited via Amazon Mechanical Turk.[5] They were first given concise definitions of all factors, then shown image-message pairs and asked to rate the influence of each factor using a 5-point Likert scale (1: not at all influential, 5: extremely influential). Each item was annotated by three annotators, with strict quality controls to remove low-quality responses.

**Results.** If VPFs capture meaningful dimensions of persuasive evidence, they should be perceived as more relevant in High items, where visual evidence more effectively supports the target message. Consistent with this expectation, High items receive significantly higher factor-importance ratings than Low items across all nine factors (rank-based AUC = 0.649–0.805, all $q < 10^{-7}$), supporting the construct validity of the VPF taxonomy. Because the validation survey and VLM evaluation differ in task format and available context, we avoid direct one-to-one comparisons between survey ratings and VLM predictions. Full details are in Appendix C.

[5]https://www.mturk.com

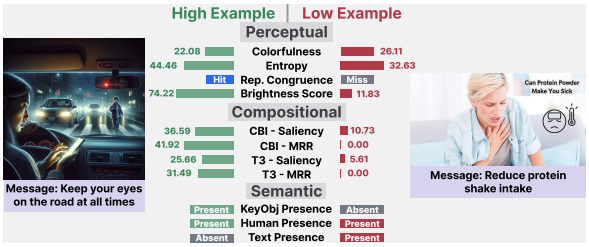



Figure 2: Qualitative comparison of VPF patterns between High- and Low-persuasiveness cases.

## 4 Analysis of Visual Persuasive Factors

Our main human–VLM comparison results are shown in Tab. 1 with qualitative examples in Fig. 2. To facilitate interpretation of Tab. 1, we report stars[6] and standard effect sizes (Cohen, 1988) for how strongly each factor separates High- from Low-persuasiveness groups: large$_l$ ($\geq 0.8$), medium$_m$ ($\geq 0.5$), small$_s$ ($\geq 0.2$), and negligible$_n$ ($< 0.2$).

**VPFs differentiate High- and Low-persuasiveness cases along interpretable visual and semantic dimensions.** To test whether VPFs capture factors that distinguish human judgments, we first examine Panel A of Tab. 1. Panel A shows that the proposed VPFs meaningfully separate human-labeled High and Low items: High items show stronger VPF signals across perceptual, compositional, and semantic levels—*e.g.*, representative color congruence, T3 saliency, and key object presence—whereas cues such as colorfulness and text presence are more associated with Low items.

**VLMs partially follow, but do not fully reproduce, human-aligned VPF patterns.** We compare Panel A and Panel B of Tab. 1. Panel B shows that VLM predictions broadly track part of this pattern: for example, representative color congruence and key object presence remain consistently associated with predicted High labels across models, with small-to-medium effect sizes. However, this alignment is only partial. Several factors, including human presence, CBI saliency, T3 saliency, and brightness score under negative strategies, show stronger associations with VLM predictions than with human judgments. The brightness results provide a useful side observation: while raw brightness is not informative, the brightness score under negative strategies shows meaningful associations, suggesting that the persuasion strategy is also a meaningful factor to consider in visual persuasive

[6]Stars indicate $p$-values: $^{***}$ $p < .001$, $^{**}$ $p < .01$, $^{*}$ $p < .05$ (two-sided).

| Group | Row | Perceptual (%) | | | | | | Compositional (%) | | | | Semantic (OR) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Color | | | Brightness | | | CBI | | T3 | | KeyObj | Human | Text |
| | | Colorfulness | Entropy | Rep. Congruence | Raw | Pos | Neg | Saliency | MRR | Saliency | MRR | | | |
| **Panel A: Ground Truth** | | | | | | | | | | | | | | |
| Human | $\Delta_f^{(Y)}$ | -3.94$^{**}{}_s$ | 4.57$^{***}{}_s$ | 15.09$^{***}{}_s$ | -0.34$_n$ | -5.11$_s$ | 5.91$^{**}{}_s$ | 2.65$^{**}{}_s$ | -2.83$_n$ | 3.19$^{***}{}_s$ | 3.31$^{**}{}_s$ | 3.36$^{***}{}_m$ | 1.25$_n$ | 0.60$^{*}{}_s$ |
| **Panel B: VLM Predictions** | | | | | | | | | | | | | | |
| GPT-5 | $\Delta_f^{(\hat{Y})}$ | -1.66$_n$ | 4.82$^{***}{}_s$ | 12.34$^{***}{}_s$ | -1.69$_n$ | -5.74$^{*}{}_s$ | 4.19$^{*}{}_s$ | 3.27$^{***}{}_s$ | 2.45$_n$ | 2.58$^{***}{}_s$ | 1.23$_n$ | 2.87$^{***}{}_m$ | 1.56$^{*}{}_s$ | 0.94$_n$ |
| GPT-5-mini | $\Delta_f^{(\hat{Y})}$ | -3.51$^{*}{}_s$ | -3.68$^{**}{}_s$ | 16.19$^{***}{}_s$ | -1.84$_n$ | -6.92$^{*}{}_s$ | 6.67$^{**}{}_s$ | 2.90$^{***}{}_s$ | 1.46$_n$ | 2.65$^{***}{}_s$ | 0.67$_n$ | 2.68$^{***}{}_m$ | 1.58$^{*}{}_s$ | 1.35$_n$ |
| Gemma3 | $\Delta_f^{(\hat{Y})}$ | -2.87$_n$ | 3.08$^{**}{}_s$ | 15.04$^{***}{}_s$ | -0.67$_n$ | -6.30$^{*}{}_s$ | 7.99$^{***}{}_m$ | 7.01$^{***}{}_m$ | 4.73$^{*}{}_s$ | 4.97$^{***}{}_m$ | 2.78$^{*}{}_s$ | 2.03$^{***}{}_s$ | 1.88$^{**}{}_s$ | 1.37$_n$ |
| Qwen2.5-VL | $\Delta_f^{(\hat{Y})}$ | -2.51$_n$ | 4.99$^{***}{}_s$ | 14.41$^{**}{}_s$ | 0.86$_n$ | -5.35$_s$ | 8.78$^{***}{}_m$ | 6.15$^{***}{}_m$ | 1.60$_n$ | 5.49$^{***}{}_l$ | 4.73$^{***}{}_s$ | 2.73$^{***}{}_m$ | 2.48$^{***}{}_m$ | 1.47$_s$ |
| **Panel C: VLM False Positives among Ground-Truth Low Items** | | | | | | | | | | | | | | |
| GPT-5 | $\Delta_f^{(FP-TN)}$ | 0.31$_n$ | 3.28$^{*}{}_s$ | 4.60$_n$ | -2.83$_n$ | -6.10$_s$ | 0.51$_n$ | 0.42$_n$ | 0.26$_n$ | 1.31$_n$ | -0.66$_n$ | 2.18$^{***}{}_s$ | 1.72$^{*}{}_s$ | 1.19$_n$ |
| GPT-5-mini | $\Delta_f^{(FP-TN)}$ | -2.05$_n$ | 2.29$_n$ | 7.80$_n$ | -2.53$_n$ | -5.99$_s$ | 3.81$_s$ | 1.06$_n$ | -0.90$_n$ | 1.78$^{*}{}_s$ | 1.28$_n$ | 2.02$^{**}{}_s$ | 1.53$_s$ | 1.79$^{*}{}_s$ |
| Gemma3 | $\Delta_f^{(FP-TN)}$ | -0.93$_n$ | 2.90$_s$ | 6.00$_n$ | -1.90$_n$ | -6.69$^{*}{}_s$ | 5.99$^{*}{}_s$ | 2.23$^{*}{}_s$ | 1.64$_n$ | 2.47$^{**}{}_s$ | 0.85$_n$ | 1.41$_n$ | 1.96$^{**}{}_s$ | 1.63$^{*}{}_s$ |
| Qwen2.5-VL | $\Delta_f^{(FP-TN)}$ | -1.73$_n$ | 4.20$^{**}{}_s$ | 14.60$^{**}{}_s$ | 0.31$_n$ | -5.35$_s$ | 8.08$^{**}{}_s$ | 3.60$^{**}{}_s$ | 0.49$_n$ | 3.42$^{***}{}_m$ | -0.21$_n$ | 1.99$^{**}{}_s$ | 2.17$^{***}{}_s$ | 1.56$_n$ |

Table 1: Human–VLM comparison summary across VPFs. Panel A shows that VPFs separate human-labeled High and Low items, while Panel B shows that VLM predictions partially reproduce this human-aligned pattern. Panel C further analyzes VLM false positives among ground-truth Low items, revealing that models often associate persuasion-relevant cues with erroneous High predictions. **Overall, VLMs appear sensitive to several persuasive cues, but their false positives suggest an overgeneralization from persuasive cues to High persuasiveness.**

understanding. Conversely, text presence is associated with human-labeled Low persuasiveness but tends to be neutral or High for VLMs, and colorfulness shows a negative human association that models reflect only weakly and inconsistently. Overall, VLMs appear sensitive to several human-aligned VPFs, but they do not consistently weight or contextualize these factors in the same way as humans.

**VLM false positives arise from overgeneralizing persuasion-relevant cues.** Motivated by the recall-oriented bias observed in §3.1, we further analyze VLM false positives through the lens of VPFs. Panel C of Tab. 1 focuses on human-labeled Low items and compares cases that VLMs incorrectly predict as High against those correctly predicted as Low. This comparison reveals that some factors that look human-aligned in Panel B can also drive erroneous High predictions. For example, key object presence is associated with human High labels and VLM High predictions in Panels A and B, but it is also significantly associated with false positives in Panel C. Moreover, human and text presence are more strongly tied to VLM false positives than to human persuasiveness judgments, suggesting that models overgeneralize semantic presence cues as persuasive evidence. A similar trend also appears for some perceptual and compositional cues, with factors such as T3 saliency, color entropy, and brightness score under negative strategies elevated among false positives. Since all items in this comparison are human-labeled Low, these results indicate that VLMs often mistake plausible persuasive visual cues for sufficient evidence of persuasiveness, rather than evaluating whether those cues genuinely support the target message.

## 5 VPF-Guided Knowledge Injection and Reasoning

The analysis in §4 motivates two questions. First, if VLMs are already sensitive to VPFs but overuse them as generic positive signals, can explicit VPF knowledge help models treat these cues as context-dependent evidence rather than sufficient evidence of persuasiveness? Second, can reasoning scaffolds help models integrate available visual evidence into a persuasive judgment? We refer to the latter ability as *persuasive integration*: the ability to use visual cues as evidence for or against the communicative intent of an image-message pair. We test this via VPF-guided interventions, focusing on key object cues because they reliably separate human judgments yet also contribute to VLM false positives.

### 5.1 Task Description

**Testing cue use through knowledge injection.** To test whether explicit VPF knowledge helps models avoid treating persuasion-relevant cues as determinative of persuasiveness, we consider two injection variants: Factor-aware Cognitive Injection and Aligned Key Object Context. (i) Factor-aware Cognitive Injection provides the model with explicit knowledge of the proposed VPFs during evaluation, either one factor at a time or all nine factors together, while presenting them only as

| Non-Aware Key Object | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Baseline** | | | | **Chain-of-Thoughts** | | | | **Knowledge-Conditioned Chain** | | | |
| | Acc. | Prec. | Rec. | F1 | Acc. | Prec. (Δ) | Rec. (Δ) | F1 (Δ) | Acc. | Prec. (Δ) | Rec. (Δ) | F1 (Δ) |
| GPT-5 | 69.57 | 51.01 | 98.88 | 67.30 | 70.11 | 51.52 ↑0.51 | 94.94 ↓3.94 | 66.80 ↓0.50 | 65.48 | 47.81 ↓3.20 | 98.31 ↓0.57 | 64.34 ↓2.96 |
| GPT-5-mini | 60.14 | 44.13 | 97.19 | 60.70 | 63.88 | 46.63 ↑2.50 | 97.19 (0.00) | 63.02 ↑2.32 | 63.35 | 46.30 ↑2.17 | 98.31 ↑1.12 | 62.95 ↑2.25 |
| Gemma3 | 63.88 | 46.33 | 88.76 | 60.89 | 51.25 | 38.84 ↓7.49 | 93.82 ↑5.06 | 54.93 ↓5.96 | 39.15 | 33.98 ↓12.35 | 97.75 ↑8.99 | 50.43 ↓10.46 |
| Qwen2.5-VL | 54.98 | 40.51 | 89.89 | 55.85 | 68.09 | 49.82 ↑9.31 | 76.40 ↓13.49 | 60.31 ↑4.46 | 57.65 | 42.46 ↑1.95 | 94.94 ↑5.05 | 58.68 ↑2.83 |
| **Aware Key Object** | | | | | | | | | | | | |
| **Model** | **Aligned Key Object Context** | | | | **Key Object-Aware CoT** | | | | **Rationale Injection** | | | |
| | Acc. | Prec. (Δ) | Rec. (Δ) | F1 (Δ) | Acc. | Prec. (Δ) | Rec. (Δ) | F1 (Δ) | Acc. | Prec. (Δ) | Rec. (Δ) | F1 (Δ) |
| GPT-5 | 69.93 | 51.44 ↑0.43 | 90.45 ↓8.43 | 65.58 ↓1.72 | 63.70 | 46.41 ↓4.60 | 94.38 ↓4.50 | 62.22 ↓5.08 | 69.40 | 50.89 ↓0.12 | 96.07 ↓2.81 | 66.54 ↓0.86 |
| GPT-5-mini | 58.72 | 43.11 ↓1.02 | 94.94 ↓2.25 | 59.30 ↓1.40 | 54.98 | 41.18 ↓2.95 | 98.31 ↑1.12 | 58.04 ↓2.66 | 64.06 | 46.79 ↑2.66 | 98.31 ↑1.12 | 63.41 ↑2.71 |
| Gemma3 | 62.28 | 45.06 ↓1.27 | 87.08 ↓1.68 | 59.39 ↓1.50 | 46.44 | 36.89 ↓9.44 | 97.19 ↑8.43 | 53.48 ↓7.41 | 59.79 | 43.65 ↓2.68 | 92.70 ↑3.94 | 59.35 ↓1.54 |
| Qwen2.5-VL | 53.02 | 39.56 ↓0.95 | 91.57 ↑1.68 | 55.25 ↓0.60 | 62.46 | 45.22 ↑4.71 | 87.64 ↓2.25 | 59.66 ↑3.81 | 53.91 | 39.37 ↓1.14 | 84.27 ↓5.62 | 53.67 ↓2.18 |

Table 2: Results of semantic grounding and reasoning scaffolds on VLM performance. Δ next to Precision, Recall, and F1 shows the absolute percentage-point change relative to the model's own baseline scores (↑ / ↓ for direction).

| Injection Type | ΔFP Rate | ΔPrec. | ΔRec. | ΔF1 | No. of Models with F1 Gain |
|---|---|---|---|---|---|
| **Panel A. Factor-wise Cognitive Injection** | | | | | |
| Color | ↓6.32 | ↑2.58 | ↓2.39 | ↑1.62 | 3/4 |
| Brightness | ↓5.01 | ↑2.15 | **↓0.70** | ↑1.76 | 3/4 |
| Saliency-Center | ↓5.66 | ↑2.10 | ↓2.39 | ↑1.30 | 3/4 |
| MRR-Center | ↓5.34 | ↑2.12 | ↓1.54 | ↑1.53 | 3/4 |
| Saliency-RoT | ↓7.68 | ↑3.25 | ↓2.25 | **↑2.32** | 4/4 |
| MRR-RoT | ↓5.08 | ↑1.94 | ↓1.54 | ↑1.38 | 3/4 |
| KeyObj | ↓3.91 | ↑1.46 | ↓1.69 | ↑0.88 | 4/4 |
| Human | **↓9.38** | **↑3.61** | ↓4.35 | ↑1.95 | 4/4 |
| Text | ↓5.27 | ↑2.13 | ↓1.40 | ↑1.52 | 3/4 |
| **Panel B. All-factor Cognitive Injection** | | | | | |
| All | ↓1.03 | ↑0.42 | ↓0.28 | ↑0.31 | 2/4 |

Table 3: Average false-positive mitigation and performance effects of cognitive injection across four VLMs relative to each model's baseline in Tab. 2. All Δ values are computed as injection − baseline.

context-dependent auxiliary cues and not prescribing any fixed effect on the persuasive judgment. (ii) Aligned Key Object Context supplies compact structured information about which detected key object candidates are semantically aligned with the message (*e.g.*, "devices" for the message "remove unnecessary electronic devices"). If no key object is found, no additional context is provided.

**Testing persuasive integration through reasoning scaffolds.** To examine whether structured reasoning can improve persuasive judgment, we consider three scaffold variants: (i) general Chain-of-Thought (CoT), which appends the instruction "*Let's think step by step*" to encourage explicit stepwise reasoning (Wei et al., 2022); (ii) Knowledge-Conditioned Chain, a lightweight reasoning scaffold that explicitly prompts the model to (a) identify candidate key objects, (b) assess their alignment with the message, and (c) make a final persuasiveness judgment; and (iii) Key Object-Aware CoT, which uses the same step-by-step prompting while additionally informing the model of the relevant key object cue (*i.e.*, Aligned Key Object Context setting). Both task prompts are in Appendix D.

## 5.2 Results

**Explicit VPF knowledge reveals cue-framing sensitivity.** In Tab. 3, Factor-aware Cognitive Injection yields a general trend of lower false-positive rates and higher F1 scores across models for most VPFs (full results are in Appendix D.2). This suggests that making VPFs explicit as context-dependent supporting cues can mitigate the recall-oriented bias by discouraging models from treating persuasion-relevant cues as sufficient evidence of persuasiveness. Thus, the bias is not merely a fixed tendency; VLM judgments are sensitive to how persuasive cues are framed. The contrast with Aligned Key Object Context in Tab. 2 makes this clearer. The consistent F1 drops show that direct key object context does not resolve the bias, aligning with the earlier finding that key object cues are associated with VLM false positives. This indicates that direct cue specification, without contextual framing, is insufficient; VLMs still lack a stable criterion for when such cues should count as persuasive evidence.

**Reasoning scaffolds expose the integration bottleneck.** Reasoning-based experiments show that explicit self-reasoning is not, by itself, a robust remedy for persuasive judgment. General CoT and Knowledge-Conditioned Chain yield mixed effects, indicating that simply prompting the model to reason step by step does not consistently improve performance. More importantly, when VLMs reason with explicit key object cues, the recall-oriented bias can be amplified, as shown by Key Object-Aware CoT (*e.g.*, Gemma3: Prec. −9.44, Rec.

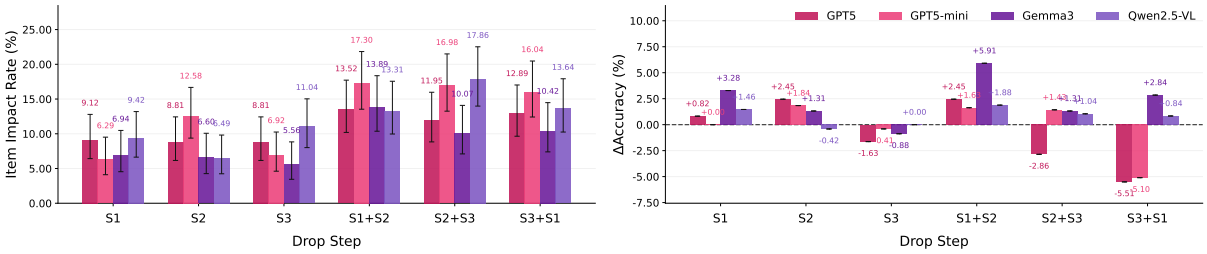



Figure 3: (Left) Item impact rate, the proportion of predictions changed by each single- or pair-step ablation. (Right) Accuracy change, measuring the corresponding change in binary prediction accuracy.

+8.43, F1 −7.41). Together with the knowledge injection results, this suggests that the remaining bottleneck lies in persuasive integration. This motivates a finer-grained diagnosis of where the integration process breaks down.

### 5.3 Diagnosing Persuasive Integration through Rationale Analysis

We analyze model rationales by decomposing them into functional steps to identify which components are most critical for persuasive integration. Prior work treats rationales as supporting evidence for model predictions and evaluates their alignment and faithfulness (Lei et al., 2016; DeYoung et al., 2020). Building on this view, we use rationales as a diagnostic proxy for examining how models use evidence and integrate it into persuasive judgments.

**Step-wise rationale ablation.** For each detected key object from the VPF pipeline (§3.3), we prompt the model to generate a per-object rationale explaining whether and how that object contributes to persuasiveness (details in Appendix D.3). We decompose each rationale into three functional steps: $S1$, which identifies the relevant object; $S2$, which explains how that evidence semantically supports or undermines the message; and $S3$, which connects the preceding analysis to the final persuasive judgment. The functional spans are tagged by GPT-5 under a text-preserving constraint. To capture not only the contribution of individual steps but also their interactions, we then perform both single- and pair-step ablations by selectively removing one or two tagged spans, and re-evaluate the model on the same item relative to the Rationale Injection output in Tab. 2. For each ablation, we measure (i) the item impact rate, defined as the rate at which predictions change, and (ii) the change in prediction accuracy against the ground-truth label. Full parsing procedure and examples are in Appendix D.4.

**Results.** Fig. 3 (Left) shows that all three rationale steps meaningfully affect model behavior. Removing a single step changes predictions for a non-trivial portion of samples, and pair-step ablations all produce double-digit item impact rates across models. Thus, earlier steps are not simply ignored; rather, each part of the rationale actively intervenes in the model's decision process. However, Fig. 3 (Right) reveals an asymmetry in how these components affect correctness. While ablating $S1$ or $S2$ often leaves accuracy unchanged or even improves it, removing $S3$ alone tends to hurt performance more directly. The clearest pattern appears in pair-step ablations: dropping $S1{+}S2$ improves accuracy across all models. This sharpens the integration-failure account. A central weakness appears to lie in how object identification ($S1$) and semantic message alignment ($S2$) are connected to the final persuasive claim. In other words, VLMs do not reliably know how that evidence should count for or against the communicative intent of the image-message pair. We report an additional false-positive-focused analysis in Appendix D.4.

## 6 Conclusion

Our work investigates visual persuasive understanding in VLMs from a cognitive psychology perspective. We introduce Visual Persuasive Factors, a structured taxonomy for analyzing how visual cues shape persuasiveness judgments. We show that even in high-consensus cases, VLMs exhibit a recall-oriented bias; VPF analysis further reveals that VLMs only partially reproduce human-aligned factor patterns, and their false positives arise from conflating persuasion-relevant cues with message support. Building on this insight, we evaluate VPF-guided interventions and find that explicit factor knowledge can help only when framed as auxiliary cues rather than fixed evidence, while naive cue grounding or step-by-step reasoning is insufficient. Finally, rationale ablation identifies persuasive integration as a key bottleneck, highlighting that VLM-based visual persuasion evaluation requires connecting visual evidence to communicative intent.

## Limitations

The primary objective of this study is to establish a foundational diagnostic framework for assessing the interpretive capabilities of VLMs regarding visual persuasiveness through the proposed VPFs. In light of this goal, we acknowledge several constraints in our current scope. First, by focusing on high-agreement scenarios, our study did not account for individual personality traits, which can significantly influence perceptions of persuasion. A comprehensive understanding would require integrating these personal factors into the analysis. Future research will therefore aim to integrate the proposed VPF framework with user-centric modeling and psychographic profiles to investigate how VLMs can navigate the inherent subjectivity and individual variance in persuasive communication. Second, our VPF analysis primarily examines factors at the level of individual perceptual, compositional, and semantic cues. However, visual persuasion can emerge from interactions among multiple factors, such as how object semantics, spatial composition, color, and message framing jointly shape judgment. Future work will extend VPFs beyond level-wise analysis toward modeling cross-factor interactions and combined cue configurations.

## Ethical Considerations

We acknowledge the dual-use concerns inherent in persuasion research. A deeper understanding of the factors driving visual persuasion could potentially be misused to create more effective disinformation or manipulative propaganda. However, we also believe this research has significant positive societal potential, such as in designing more effective public health and safety communications (*e.g.*, "always turn off the stove"). Our contribution is focused on the fundamental analysis of VLMs' cognitive alignment and its current limitations, not on building persuasive applications. We hope this research provides a foundation for developing VLMs that can better understand complex human cognitive processes in a responsible manner.

Furthermore, while we report effect sizes based on standard criteria for formal consistency (Cohen, 1988), we emphasize that these effects can have a much larger impact in real-world settings than the reported metrics suggest. As argued in the field of psychology by Funder and Ozer (2019), effects typically categorized as "small" by conventional standards can be highly consequential when they operate repeatedly, affect large populations, or accumulate over time. In real-world visual persuasion settings, even small effects may become consequential when repeatedly deployed at scale. This motivates caution in interpreting and applying VPF-based analyses, especially in systems that could be used to generate or optimize persuasive visual content.

## References


Francis M. Adams and Charles E. Osgood. 1973. A cross-cultural study of the affective meanings of color. *Journal of Cross-Cultural Psychology*, 4(2):135–156.

Arjun R Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T Freeman, and 1 others. 2023. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23201–23211.

Katherine Armstrong, Adam S. Richards, and Kylie J. Boyd. 2021. Red-hot reactance: Color cues moderate the freedom threatening characteristics of health psas. *Health Communication*, 36(6):663–670. PMID: 31818126.

H. Bai, J. G. Voelkel, S. Muldowney, and 1 others. 2025a. Llm-generated messages can persuade humans on policy issues. *Nature Communications*, 16:6037.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Noel T. Brewer, Marissa G. Hall, Seth M. Noar, Humberto Parada, Annette Stein-Seroussi, Laura E. Bach, Sean Hanley, and Kurt M. Ribisl. 2016. Effect of pictorial cigarette pack warnings on changes in smoking behavior: A randomized clinical trial. *JAMA Internal Medicine*, 176(7):905–912.

Daniel Chandler and Rod Munday. 2011. *A Dictionary of Media and Communication*. Oxford University Press, USA. EISBN: 9780191727979.

Jiwan Chung, Sungjae Lee, Minseo Kim, Seungju Han, Ashkan Yousefpour, Jack Hessel, and Youngjae Yu. 2024. Selective vision is the challenge for visual reasoning: A benchmark for visual argument understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2423–2451, Miami, Florida, USA. Association for Computational Linguistics.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Routledge.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. Measuring the persuasiveness of language models.

Ryan S. Elder and Aradhna Krishna. 2012. The "visual depiction effect" in advertising: Facilitating embodied mental simulation through product orientation. *Journal of Consumer Research*, 38(6):988–1003.

David C. Funder and Daniel J. Ozer. 2019. Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2):156–168.

Karl R Gegenfurtner and Jochem Rieger. 2000. Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology*, 10(13):805–808.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

David Hasler and Sabine E. Suesstrunk. 2003. Measuring colorfulness in natural images. In *Human Vision and Electronic Imaging VIII*, volume 5007, pages 87 – 95. International Society for Optics and Photonics, SPIE.

Chris Janiszewski. 1990. The influence of print advertisement organization on affect toward a brand name. *Journal of Consumer Research*, 17(1):53–65.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. Ultralytics yolov8. `https://github.com/ultralytics/ultralytics`.

Jungseock Joo, Weixin Li, Francis F. Steen, and Song-Chun Zhu. 2014. Visual Persuasion: Inferring Communicative Intents of Images . In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 216–223, Los Alamitos, CA, USA. IEEE Computer Society.

Raymond P. Voss Jr., Ryan Corser, Michael McCormick, and John D. Jasper. 2018. Influencing health decision-making: A study of colour and message framing. *Psychology & Health*, 33(7):941–954. PMID: 29667448.

Junseo Kim, Jongwook Han, Dongmin Choi, Jongwook Yoon, Eun-Ju Lee, and Yohan Jo. 2025. PVP: An image dataset for personalized visual persuasion with persuasion strategies, viewer characteristics, and persuasiveness ratings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19209–19237, Vienna, Austria. Association for Computational Linguistics.

Bert P. Krages. 2005. *The Art of Composition: A Guide to Creative Photography*. Allworth Communications, Inc., New York, NY.

Yaman Kumar, Rajat Jha, Arunim Gupta, Milan Aggarwal, Aditya Garg, Tushar Malyan, Ayush Bhardwaj, Rajiv Ratn Shah, Balaji Krishnamurthy, and Changyou Chen. 2023. Persuasion strategies in advertisements. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):57–66.

Lauren Labrecque and George Milne. 2011. Exciting red and competent blue: The importance of color in marketing. *Journal of The Academy of Marketing Science - J ACAD MARK SCI*, 40.

Lauren I. Labrecque, Vanessa M. Patrick, and George R. Milne. 2013. The marketers' prismatic palette: A review of color research and future directions. *Psychology & Marketing*, 30(2):187–202.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. 2021. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 12919–12928.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Long Mai, Hoang Le, Yuzhen Niu, and Feng Liu. 2011. Rule of thirds detection from photograph. In *2011 IEEE International Symposium on Multimedia*, pages 91–96.

Daniela Mapelli and Marlene Behrmann. 1997. The role of color in object recognition: Evidence from visual agnosia. *Neurocase*, 3(4):237–247.

S. C. Matz, J. D. Teeny, S. S. Vaid, and 1 others. 2024. The potential of generative ai for personalized persuasion at scale. *Scientific Reports*, 14:4692.

Lizzie Maughan, Sergei Gutnikov, and Rob Stevens. 2007. Like more, look more. look more, like more: The evidence from eye-tracking. *Journal of Brand Management*, 14(4):335–342.

K. McLAREN. 1976. Xiii—the development of the cie 1976 (l* a* b*) uniform colour space and colour-difference formula. *Journal of the Society of Dyers and Colourists*, 92(9):338–341.

Lu (Monroe) Meng, Sining Kou, Shen Duan, Yijie Wang, and Kevin Lü. 2026. Effect of human images in advertisements on consumers' experiential purchase intentions. *Journal of Advertising*, 55(2):184–204.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. 2022. Simple open-vocabulary object detection with vision transformers. In *European Conference on Computer Vision (ECCV)*.

Seth M Noar, Marissa G Hall, Diane B Francis, Kurt M Ribisl, Jessica K Pepper, and Noel T Brewer. 2016. Pictorial cigarette pack warnings: a meta-analysis of experimental studies. *Tobacco Control*, 25(3):341–354.

OpenAI. 2025. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf.

Richard E. Petty and John T. Cacioppo. 1986. *The Elaboration Likelihood Model of Persuasion*, pages 1–24. Springer New York, New York, NY.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *Preprint*, arXiv:2401.14159.

Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Empathetic persuasion: Reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 844–856, Seattle, United States. Association for Computational Linguistics.

Matthew Selove. 2014. A dynamic model of competitive entry response. *Marketing Science*, 33(3):353–363.

Kiwon Seo. 2020. Meta-analysis on visual persuasion– does adding images to texts influence persuasion? *ATHENS JOURNAL OF MASS MEDIA AND COMMUNICATIONS*, 6:177–190.

Somesh Singh, Yaman K Singla, Harini SI, and Balaji Krishnamurthy. 2024. Measuring and improving persuasiveness of large language models. *Preprint*, arXiv:2410.02653.

Benjamin W. Tatler. 2007. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):4–4.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. 2024. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. 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, Florence, Italy. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2024. Large language models as analogical reasoners. In *The Twelfth International Conference on Learning Representations*.

Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. 2008. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32–32.

Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. 2018. Equal but not the same: Understanding the implicit relationship between persuasive images and text. *arXiv preprint arXiv:1807.08205*.

## A Dataset Details

**Dataset and agreement filtering.** We use the PVP dataset (Kim et al., 2025), which is released under the CC-BY-4.0. Each message in the dataset is paired with ∼54 images (different images for the same message). Each image is rated by four annotators with distinct psychological profiles, who assign a persuasiveness score from 0 to 10 based on the (message, image) pair. Because four raters may agree or disagree per image, we analyze inter-annotator agreement and then filter the dataset accordingly. Specifically, we consider two schemes: (i) the original 0–10 integer scores and (ii) a banded scheme (*low*=0–2, *mid*=3–7, *high*=8–10). Fleiss' $\kappa$ score over the entire corpus is low for both schemes, but banding increases agreement and yields many "Almost perfect" items (Tab. 4).

**Robust high-consensus core subset.** To obtain a robust subset that reflects stable human cues, we select only the banded items with "Almost perfect" agreement among the four raters and keep images whose consensus label is *low* or *high* (discarding *mid*). After deduplication by image path, this yields 562 images (384 *low*, 178 *high*) out of 2,789 "Almost perfect" images, excluding 2,227 *mid*-only cases. This process supports the assumption that highly consistent human judgments capture stable persuasive signals, enabling a more direct and noise-resistant comparison with model attributions.

## B Detailed VPF and Extracted Examples

This section provides the detailed definitions and equations for the VPF features described in Section 3.3.

### B.1 Perceptual Level

The following section describes how we extracted features related to color and brightness and how we used them as comparison metrics. For metrics such as colorfulness and entropy, we reported the normalized values (*i.e.*, those included in Tab. 1) for comparison purposes.

- **Colorfulness:** To measure color intensity, we use the colorfulness metric of Hasler and Suesstrunk (2003). Given RGB channels, we first compute two opponent-color components:

$$o_{rg} = R - G, \tag{2}$$

$$o_{yb} = \frac{1}{2}(R + G) - B. \tag{3}$$

We then summarize their dispersion and average magnitude as

$$s_o = \sqrt{\mathrm{std}(o_{rg})^2 + \mathrm{std}(o_{yb})^2}, \tag{4}$$

$$m_o = \sqrt{\mathrm{mean}(o_{rg})^2 + \mathrm{mean}(o_{yb})^2}, \tag{5}$$

and define the image colorfulness score as

$$C_{\text{colorfulness}} = s_o + 0.3\, m_o. \tag{6}$$

*Interpretation*: Typical reference values are 0 (not colorful), 15 (slightly), 33 (moderately), 45 (average), 59 (quite), 82 (highly), and 109 (extremely colorful).

| | **High** | **Low** | **Effect** ($d$, **High–Low**) |
|---|---|---|---|
| Human | 29.28 | 33.22 | -0.23 |
| GPT-5 | 31.33 | 32.99 | -0.10 |
| GPT-5-mini | 30.91 | 34.42 | -0.21 |
| Gemma3 | 30.85 | 33.71 | -0.17 |
| Qwen2.5 | 31.23 | 33.74 | -0.15 |

Table 5: Normalized colorfulness scores (Hasler metric) for high- and low-persuasiveness groups, with Cohen's $d$ effect sizes.

- **Color Entropy:** To measure color diversity, we converted the image into the CIELAB color space and quantized the chrominance channels $(a^*, b^*)$ into a 2D histogram. We then calculated the amount of information in the color distribution, $H_{\text{color}}$, using the Shannon entropy formula, where a higher value indicates a more uniform distribution of diverse colors.

$$H_{\text{color}} = -\sum_i p_i \log_2 p_i \tag{7}$$

where $p_i$ is the probability of the $i$-th color bin in the $a^*b^*$ histogram.

| | **High** | **Low** | **Effect** ($d$, **High–Low**) |
|---|---|---|---|
| Human | 42.97 | 38.40 | 0.35 |
| GPT-5 | 41.71 | 36.89 | 0.37 |
| GPT-5-mini | 40.96 | 37.28 | 0.28 |
| Gemma3 | 41.06 | 37.98 | 0.24 |
| Qwen2.5 | 41.33 | 36.34 | 0.39 |

Table 6: Normalized color entropy scores for high- and low-persuasiveness groups, with Cohen's $d$ effect sizes.

- **Representative Color Congruence**: To measure the alignment between perceptual color properties and semantic framing, we define a discrete

| Metric | Original (0–10) | Banded (low/mid/high) |
|---|---|---|
| Overall Fleiss' $\kappa$ | 0.0233 | 0.0629 |
| Mean per-item $\kappa$ | 0.023 | 0.063 |
| Median per-item $\kappa$ | 0.003 | 0.011 |
| *Kappa level distribution (Landis and Koch, 1977) (count of images)* | | |
| Worse than chance | 12,424 | 13,186 |
| Slight | 13,176 | 11,318 |
| Fair | 1,264 | 1,035 |
| Moderate | 1,501 | 126 |
| Substantial | 3 | 0 |
| Almost perfect | 86 | 2,789 |
| *Robust subset (constructed from **banded** "Almost perfect")* | | |
| Almost-perfect images (banded) | — | 2,789 |
| Kept (low/high only, dedup by image) | — | 562 (low: 384, high: 178) |
| Mid-only excluded | — | 2,227 |

Table 4: Inter-annotator agreement statistics and the construction of the *robust* subset. We compare the original 0–10 integer scores (rounded) with a banded scheme that maps scores to *low* (0–2), *mid* (3–7), and *high* (8–10). Fleiss' $\kappa$ is reported overall as well as per-item distribution of agreement levels.

eight-category color palette based on the affective color space introduced by Adams and Osgood (1973). Each color is associated with a core affective profile (*e.g.*, *strong*, *pleasant*, *calm*), as summarized in Tab. 7.

| Color | Core Affective Profile | Proportion (%) |
|---|---|---|
| Red | strong, salient | 0.71 |
| Blue | pleasant, calm | 12.81 |
| Green | good, calm | 2.14 |
| Yellow | positive, active | 4.27 |
| White | pleasant, weak | 19.93 |
| Black | bad, strong | 31.67 |
| Grey | bad, passive | 17.97 |
| Brown/Purple | contextual, variable | 10.50 |

Table 7: Discrete color palette, affective profiles based on Adams and Osgood (1973), and representative color distributions extracted from dominant K-means clusters.

For the image-side representation, we extract a representative color using K-means clustering ($K = 8$). We select the dominant cluster, defined as the cluster with the largest pixel proportion, and map its centroid RGB value to the nearest palette category. The resulting distribution of representative colors is shown in Tab. 7.

Motivated by our validation survey C, where annotators selected 2.99 color categories per item on average, we infer the top-3 expected colors rather than a single color. For the message-side expectation, we use GPT-5 to infer three colors given the message and its persuasion premise, under three prompting conditions: (i) message only, (ii) message + premise, and (iii) message + premise with explicit affective profiles. We then compute *hit@3* overlap between image and message color categories. As shown in Tab. 10, incorporating both premise and affective profiles improves alignment, though effect sizes remain small to moderate. We additionally report the distribution of GPT-5 predicted color categories (iii) in Tab. 8.

| Color | High (%) | Low (%) |
|---|---|---|
| Black | 25.66 | 16.06 |
| Blue | 10.11 | 21.18 |
| Brown/Purple | 0.00 | 0.17 |
| Green | 7.68 | 14.32 |
| Grey | 15.36 | 11.37 |
| Red | 22.85 | 16.67 |
| White | 1.31 | 5.64 |
| Yellow | 17.04 | 14.58 |

Table 8: Distribution of GPT-5 predicted top-3 color categories (message+ premise + affective profile) for high- and low-persuasiveness cases.

| Color | High $\Delta$% | Low $\Delta$% |
|---|---|---|
| Red | 6.93 | 9.38 |
| Blue | -3.00 | -5.56 |
| Green | -3.93 | -6.42 |
| Yellow | -8.43 | 4.25 |
| White | -1.12 | -4.25 |
| Black | 0.19 | -0.09 |
| Grey | 9.36 | 1.30 |
| Brown/Purple | 0.00 | 1.39 |

Table 9: Ablation results using affective profiles only.

To examine whether the model is biased toward

| | Message Only | + Premise | + Premise + Affective Profile | Effect ($h$, High–Low) |
|---|---|---|---|---|
| Human (High) | 0.410 | 0.500 | 0.534 | 0.304 |
| Human (Low) | 0.333 | 0.427 | 0.383 | |
| GPT-5 | 0.332 / 0.358 | 0.458 / 0.438 | 0.478 / 0.355 | 0.251 |
| GPT-5-mini | 0.355 / 0.365 | 0.459 / 0.470 | 0.480 / 0.318 | 0.333 |
| Gemma3 | 0.375 / 0.330 | 0.463 / 0.430 | 0.490 / 0.339 | 0.307 |
| Qwen2.5 | 0.362 / 0.347 | 0.466 / 0.413 | 0.473 / 0.329 | 0.295 |

Table 10: Hit@3 overlap between image representative color and predicted colors under different prompting conditions. For VLMs, each cell reports prediction High / Low values. Effect sizes ($h$) (Cohen, 1988) correspond to the High–Low difference for the + *Premise* + *Affective Profile* condition (OptA Hit@3). Statistical significance markers are reported in Tab. 1.

affective descriptors rather than color semantics, we conduct an ablation where only affective profiles are provided without explicit color categories. As shown in Tab. 9, the resulting distribution differs substantially from the color-aware setting, indicating that the model is not merely relying on affective descriptors. Instead, meaningful alignment emerges only when explicit color information is jointly considered with affective profiles.

- **Brightness:** To quantify perceptual lightness, we convert each image from sRGB to the CIELAB color space and use the $L^*$ channel, which approximates human luminance perception (ranging from 0 to 100). The brightness metric is defined as the spatial mean of $L^*$:

$$\text{Brightness} = \frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W} L^*(x,y). \quad (8)$$

We implemented this computation using the `OpenCV` library (`cv2`), where $L^*$ is obtained by converting an image to the `LAB` color space and rescaling from the 0–255 range to the standard 0–100 scale. Higher values indicate perceptually brighter images.

Beyond measuring absolute brightness, we evaluate whether image brightness aligns with the semantic framing of the persuasion strategy inspired by Adams and Osgood (1973). For each image, we first compute its brightness as the spatial mean of the $L^*$ channel in CIELAB space and normalize it to $x \in [0, 1]$.

We then infer the polarity of the persuasion strategy (positive, negative, or neutral) from the message and its associated premise, using a rule-based mapping derived from the premise text. Given the normalized brightness $x$ and the inferred strategy polarity, we define a matching score that captures the degree of alignment between visual brightness and semantic intent:

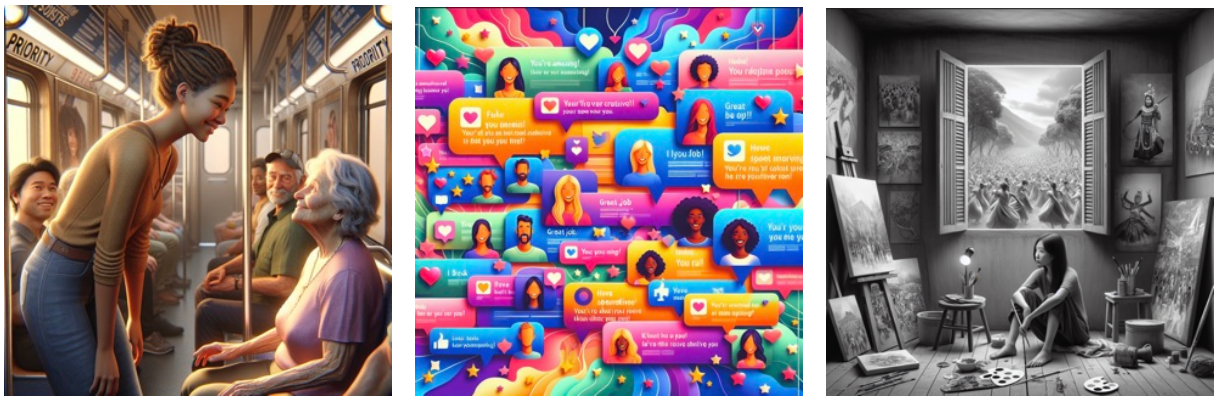
$$\text{Score}(x) = \begin{cases} x & \text{pos} \\ 1 - x & \text{neg} \\ 1 - 2|x - 0.5| & \text{neu} \end{cases} \quad (9)$$

A higher score indicates stronger alignment between image brightness and the intended persuasion framing (*e.g.*, brighter images for positive strategies and darker images for negative ones).

**Extraction Example.** We provide illustrative examples of images with varying degrees of colorfulness in Fig. 4, along with corresponding color metric values—specifically Hasler's colorfulness metric (`CF_Hasler`) and color entropy (`Entropy_bits`). We further provide representative color congruence and brightness score examples in Tab. 12.

| Example Image | `CF_Hasler` | `Entropy_bits` |
|---|---|---|
| (a) Colorfulness average | 49.37 | 6.38 |
| (b) Colorfulness high | 165.98 | 9.36 |
| (c) Colorfulness low | 1.85 | 0.95 |

Table 11: Color metrics for the examples in Fig. 4.




(a) Colorfulness average (b) Colorfulness high (c) Colorfulness low

Figure 4: Illustrative perceptual-level examples used in Appendix B.1

| Image | Clustered Palette | Message | Representative Color | Premise | GPT-5 Infer Color | Brightness $L^*$ | Brightness Score |
|---|---|---|---|---|---|---|---|
| | | Report fake news when you find it | Blue | Positive effect: Safer online environment | Top3: Blue, Green, Yellow | 73.29 | 0.733 |
| |  | Take an interest in the energy industry | Yellow | Positive effect: Others can feel inspired by your passion | Top3: Yellow, Blue, Green | 92.05 | 0.920 |
| | | Reduce protein shake intake | Black | Negative effect: Kidney strain increases significantly | Top3: Black, Red, Grey | 13.99 | 0.860 |
| | | Avoid conversations with strangers while traveling | Black | Negative effect: You can be seen as disrespectful of privacy | Top3: Black, Grey, Red | 50.21 | 0.498 |
| | | Do not visit dangerous countries | Blue | Positive effect: Others can avoid worry | Top3: Yellow, Blue, White | 54.59 | 0.546 |
| | | Clean up trash after fishing | Green | Positive effect: Others can see your care for the environment | Top3: Green, Blue, Yellow | 56.68 | 0.567 |

Table 12: Illustrative examples of representative color congruence and brightness alignment. Representative Color denotes the color category of the top cluster extracted from the clustered palette. GPT-5 Color denotes the color categories inferred to support the message. Brightness $L^*$ is the original mean lightness value in CIELAB space, and Brightness Score is computed from normalized brightness and inferred strategy polarity.

## B.2 Compositional Level

To analyze how spatial arrangement and visual hierarchy guide human attention, we utilized features based on saliency. For each image, a saliency map $S$ was extracted using the pre-trained DeepGaze IIE (Linardos et al., 2021). In addition to saliency, we construct Message-Relevant Regions (MRR) using Grounded-SAM-2 (Ren et al., 2024). We convert each message into an open-vocabulary prompt and detect candidate regions with GroundingDINO (box threshold 0.40, text threshold 0.25), followed by mask refinement using SAM2 (hiera-large). The resulting segmentation masks are used as MRR. For cases where no regions are detected under the initial prompt, we perform a fallback by re-querying with noun-based prompts extracted from the message (up to four terms) with relaxed detection thresholds (box threshold 0.25, text threshold 0.20). In the initial step, detections are obtained for $n_{\text{high}} = 130$ and $n_{\text{low}} = 247$ samples; after applying the fallback to non-detected cases, coverage increases to $n_{\text{high}} = 178$ and $n_{\text{low}} = 360$.

The following equations define the metrics:

- **Center Bias Index (CBI):** The proportion of saliency mass within a central disk. Let the image size be $H \times W$, the center $c = \left(\frac{H-1}{2}, \frac{W-1}{2}\right)$, and the radius $r_c = \alpha \min(H, W)$ with $\alpha = 0.20$.

$$\text{CBI} = \sum_{(i,j) \in D(c, r_c)} S_{i,j} \tag{10}$$

where $D(c, r) = \{(i, j) : \|(i, j) - c\|_2 \leq r\}$

- **T3:** The saliency mass within the union of four disks centered at the rule-of-thirds intersection points. Let $\mathcal{P} = \{(\frac{H}{3}, \frac{W}{3}), (\frac{H}{3}, \frac{2W}{3}), (\frac{2H}{3}, \frac{W}{3}), (\frac{2H}{3}, \frac{2W}{3})\}$ and $r_t = \beta \min(H, W)$ with $\beta = 0.10$.

$$\text{T3} = \sum_{(i,j) \in \bigcup_{p \in \mathcal{P}} D(p, r_t)} S_{i,j}. \tag{11}$$

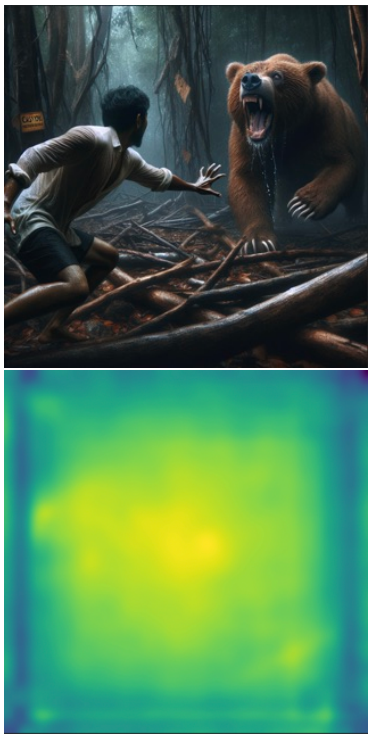
| | CBI (Saliency, %) | | T3 (Saliency, %) | |
|---|---|---|---|---|
| | High | Low | High | Low |
| Human | 34.93 | 32.28 | 20.64 | 17.45 |
| GPT-5 | 34.20 | 30.93 | 19.32 | 16.75 |
| GPT-5-mini | 33.71 | 30.82 | 19.01 | 16.36 |
| Gemma3 | 34.40 | 27.39 | 19.38 | 14.41 |
| Qwen2.5 | 33.68 | 27.53 | 18.97 | 13.49 |

Table 13: Raw compositional scores (saliency-based).

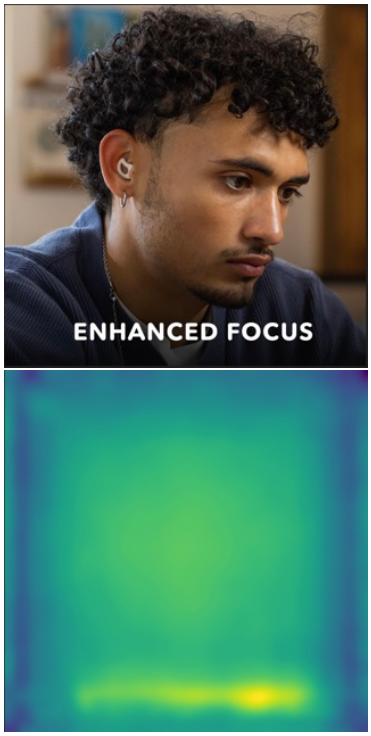
| | CBI (MRR, %) | | T3 (MRR, %) | |
|---|---|---|---|---|
| | High | Low | High | Low |
| Human | 17.27 | 20.10 | 16.81 | 13.49 |
| GPT-5 | 19.98 | 17.53 | 15.00 | 13.77 |
| GPT-5-mini | 19.46 | 18.00 | 14.73 | 14.06 |
| Gemma3 | 20.06 | 15.33 | 15.12 | 12.34 |
| Qwen2.5 | 19.31 | 17.71 | 15.03 | 10.30 |

Table 14: Raw compositional scores (MRR-based).

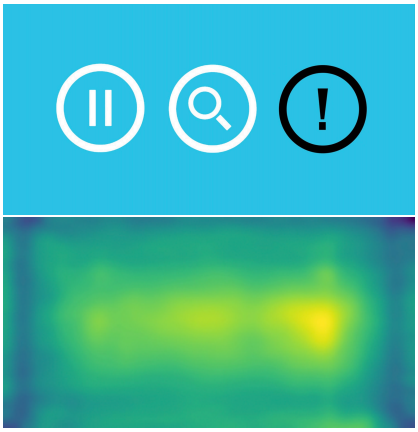
**Extraction Example.** Fig. 5 presents three representative compositional-level examples, each showing an original image (top) and its corresponding feature map (bottom). These illustrate how the model captures compositional cues—specifically saliency prediction. The high density in the saliency map means that humans will pay more attention to it. In addition, Fig. 6 shows examples of message-related region (MRR) extraction. The highlighted regions indicate segmented areas corresponding to message-relevant objects or regions, which are used to compute message-related region statistics such as center bias and rule-of-thirds overlap.



(a) Example 1




(b) Example 2



(c) Example 3

Figure 5: Illustrative examples used in Appendix B.2. Each example (top) is shown with its corresponding saliency map (bottom).

(a) Gas stove

(b) Gloves

(c) Use recyclable building materials

Figure 6: Illustrative examples of message-related region extraction used for compositional-level analysis. Highlighted regions correspond to message-relevant segmented objects or areas used to compute MRR-based composition metrics.

### B.3 Semantic Level

**Models Used.** These components collectively allow us to detect whether semantically referenced entities are visually grounded in the image. Fig. 7 summarizes the extraction pipeline for key object and human presence detection.

- **Key Noun Extraction:** We used the `en_core_web_sm` model as the parser from the `spaCy` library for Part-of-Speech (POS) tagging and noun phrase extraction. This lightweight model was selected for its balance between processing speed and linguistic accuracy.
- **Key Object Presence Detection:** We employed the OWL-ViT v1 (Minderer et al., 2022) model. OWL-ViT enables zero-shot object detection by aligning vision and text embeddings, allowing us to flexibly identify key visual concepts based on extracted textual nouns.
- **Human Presence Detection:** We used the pretrained YOLOv8-nano (Jocher et al., 2023) as the primary person detector, optimized for inference efficiency on GPU. In cases where no human instance was detected, we optionally cross-validated results using the open-vocabulary detector above.
- **Text Presence Detection:** We identify text presence using a lightweight heuristic based on the dataset-provided image descriptions. Specifically, we extract candidate textual content from quoted spans (*e.g.*, "...", '...') and ALL-CAPS phrases that resemble signage or logos, based on the observation that such patterns often correspond to explicitly rendered text in the associated images. This rule-based parsing yields 144 candidate instances out of 562 samples. Given the relatively small number, we manually verify each case to confirm whether the extracted text is visibly present in the image, resulting in 124 validated instances (22.06%).

**Metrics and Analytical Procedure.** After extracting binary indicators of key object presence ($x_{\text{obj}} \in \{0,1\}$), human presence ($x_{\text{hum}} \in \{0,1\}$) and text presence ($x_{\text{text}} \in \{0,1\}$) for each image, we constructed a feature matrix

$$X = \begin{bmatrix} 1 & x_{\text{obj}}^{(1)} & x_{\text{hum}}^{(1)} & x_{\text{text}}^{(1)} \\ 1 & x_{\text{obj}}^{(2)} & x_{\text{hum}}^{(2)} & x_{\text{text}}^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{\text{obj}}^{(n)} & x_{\text{hum}}^{(n)} & x_{\text{text}}^{(n)} \end{bmatrix},$$

| Message | Extracted KeyObj | Count |
|---|---|---|
| Be careful when using a knife | `knife` | 2 |
| Consume enough protein | `protein` | 2 |
| Wear gloves when handling chemicals | `gloves` | 1 |
| Wear a leather jacket to look stylish | `jacket` | 2 |
| Look up videos of military dog training | `dog` | 2 |
| Throw trash in the trash can | `trash` | 1 |

Table 15: Illustrative examples of key-object extraction. The extracted key object is taken from `core_noun_presence.detail.matches`; count denotes the number of detected matching regions.

where the intercept term allows for estimation of the baseline log-odds.

For each outcome variable $y \in \{0,1\}$ (*e.g.*, perceived persuasiveness or GPT agreement), we fit a logistic regression model of the form:

$$\Pr(y=1 \mid X) = \sigma(\beta_0 + \beta_1 x_{\text{obj}} + \beta_2 x_{\text{hum}} + \beta_3 x_{\text{text}})$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ denotes the logistic link function. Estimated coefficients $\hat{\beta}$ were exponentiated to yield odds ratios:

$$\text{OR}_i = e^{\hat{\beta}_i},$$

representing multiplicative changes in the odds of $y = 1$ given the presence of each feature. Robust standard errors (HC3 covariance) were used to account for potential heteroskedasticity, and all $p$-values were adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR).

This procedure allows us to quantify how semantic alignment signals—*i.e.*, the co-occurrence of mentioned and visually realized entities—contribute to persuasiveness judgments and model–human agreement.

**Extraction Example.** To illustrate how semantic-level factors are identified and mapped in our framework, Fig. 7 presents the extraction pipeline and Fig. 8 presents two representative examples showing the original images and their corresponding object-detection overlays. The extracted results for key objects and text cues are detailed in Tab. 15 and Tab. 16, respectively.

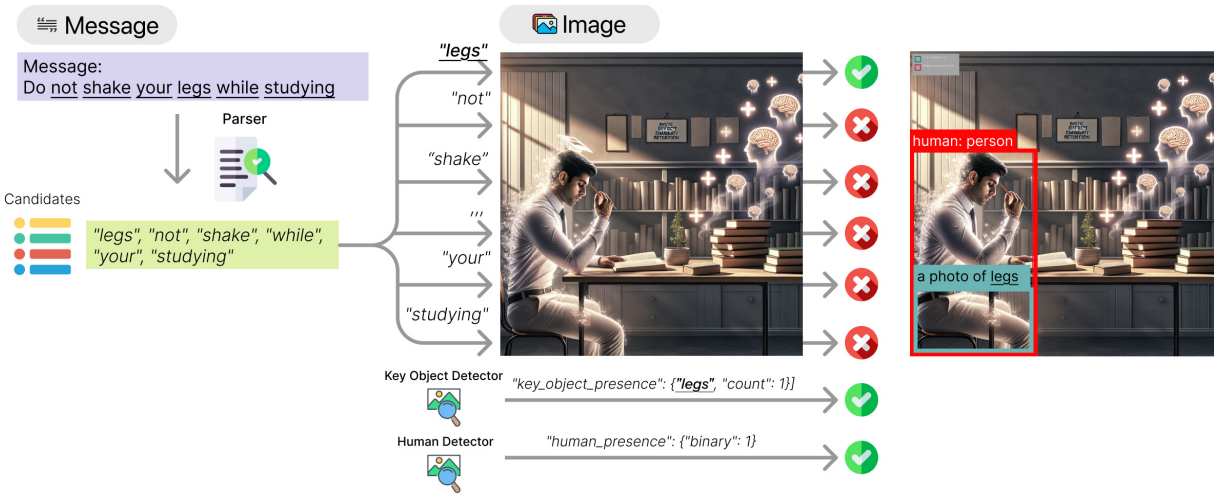



Figure 7: Overview of the semantic-level feature extraction pipeline. The pipeline detects whether semantically relevant key objects mentioned in the message are visually present in the image and identifies human presence through dedicated detectors.

(a) Example A — original

(b) Example A — detection

(c) Example B — original

(d) Example B — detection

Figure 8: Key object presence examples: each row shows a pair (original vs. object-detection overlay).

| Image | Text Mention in Image Description | Extracted Text |
|---|---|---|
| | . . . signs along the crosswalk: one on the left reading “SLOW DOWN” and another on the right reading “SLOW SAVE LIVES.” . . . | SLOW DOWN; SLOW SAVE LIVES |
| | . . . A sign to the left reads “PROTECT OUR WATERS,” emphasizing the environmental conservation theme. . . . | PROTECT OUR WATERS |
| | . . . above the toaster, there is a framed sign accompanied by the text “STAY SAFE. FEEL RESPONSIBLE.” . . . | STAY SAFE. FEEL RESPONSIBLE |
| | . . . the text on the left side of the image reads, “Does excess PROTEIN damage KIDNEYS?” . . . | Does excess PROTEIN damage KIDNEYS? |

Table 16: Illustrative examples of relevant text extraction. The extracted text corresponds to visible slogan-like or message-relevant text in the image, rather than incidental brand or location names.

## C Validation Survey Details

**Annotators.** We recruited adult annotators via Amazon Mechanical Turk (MTurk)[7] and conducted the survey through a custom web-based interface. The HIT title warned workers that the task could contain adult or sensitive content, and the task description described it as an academic study on message–image analysis and visual persuasion. Workers were required to have more than 10,000 approved HITs and an approval rate above 98%. Each HIT contained 12 image-message items and was designed to take approximately 12–15 minutes based on repeated pilot runs of the task interface. Workers were compensated at $2 per HIT, corresponding to an estimated hourly rate of approximately $8–$10. We did not collect protected demographic attributes beyond platform-provided worker identifiers used for task administration and quality control. The entire data collection process was conducted anonymously to protect participants' privacy; no personally identifiable information (PII) was collected or stored. In total, 476 participants were collected, of which 141 were retained after quality filtering.

**Task Design.** Each trial consisted of an image-only stage followed by an image-message stage. Participants were first asked to provide a binary persuasiveness judgment (Low/High) based only on the image, using a slightly adapted version of the original dataset prompt to match the binary format. This question was included to assess alignment with the original dataset judgments, yielding approximately 80% consistency. The instruction pages shown to annotators are provided in Fig. 10 and 11, while the two-stage task interface is illustrated in Fig. 12 and 13.

Subsequently, annotators rated the influence of each Visual Persuasive Factor (VPF) on a 5-point Likert scale (1: not at all influential, 2: slightly influential, 3: moderately influential, 4: very influential, 5: extremely influential). For quality check, when semantic-level factors (*e.g.*, key objects, humans, or text) were present, annotators were additionally asked to provide brief explanations (1–2 sentences) describing whether and how these elements affected their judgment. Finally, annotators provided a second binary persuasiveness judgment based on the full image-message pair.

[7] https://www.mturk.com

| Level | Factor | Human-rated Importance AUC |
|---|---|---|
| Perceptual | Color | 0.724 |
| Perceptual | Brightness | 0.654 |
| Compositional | CBI (Saliency) | 0.748 |
| Compositional | CBI (MRR) | 0.784 |
| Compositional | T3 (Saliency) | 0.744 |
| Compositional | T3 (MRR) | 0.805 |
| Semantic | Key Object | 0.755 |
| Semantic | Human | 0.709 |
| Semantic | Text | 0.649 |

Table 17: Validation survey results for the proposed visual persuasive factors. AUC denotes rank-based separation between High and Low items in human-rated factor importance.

**Annotation Protocol.** Each item was annotated by three independent participants. The full dataset of 562 image-message pairs was covered three times, resulting in triply annotated labels for all items. To prevent redundancy and learning effects, participants were not allowed to re-annotate items they had previously seen, even if they participated in multiple survey batches. As illustrated in Fig. 12 and 13, the task followed a fixed sequence within each trial: (Q1) image-only persuasiveness judgment, followed by (Q2) factor-level ratings, and (Q3) image-message persuasiveness judgment.

**Quality Control.** We implemented strict quality control procedures to ensure annotation reliability. As described in the instruction interface (Fig. 10 and 11), each task included 2 hidden attention-check (trap) questions among the 12 items; responses from annotators who failed both checks were excluded from analysis. Participants were informed of the presence of such checks but not their locations.

In addition, all free-text responses were manually reviewed by the authors, and annotations deemed uninformative or invalid were removed through a conservative filtering process. Annotators who failed attention checks or whose responses were rejected were prevented from re-participating in subsequent survey batches. Overall, this process reduced the dataset from 476 initial participants to 141 valid annotators, yielding a high-quality subset for analysis.

**Result Details.** As shown in Tab. 17, all nine factors show above-chance rank-based separation between High and Low items, with AUC values ranging from 0.649 to 0.805. Fig. 9 further visualizes the factor-wise rating distributions, showing consistently higher influence ratings for High than

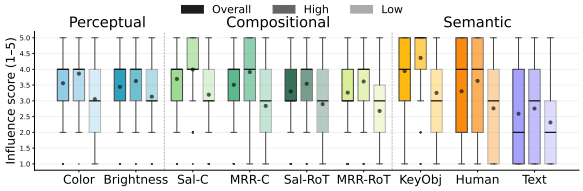



Figure 9: Factors influence ratings by 141 annotators.

Low items. Separately, under a substantially different annotation setting, new annotators' binary persuasiveness judgments align with the original GT labels at an average rate of 77.81%, suggesting that the high-consensus labels remain reasonably stable.

# Instructions

### What this study is about

You will evaluate a **image-message pair** related to **Visual Persuasion** Understanding . Visual Persuasion refers to the use of images and visual elements to influence cognition, emotion, and behavior, and it plays a central role in advertising, political communication, and social media.

In this study, we aim to understand **how much different 'Visual Persuasive Factors' influence your overall impression of persuasiveness** for the given image–message pair.

### What you will do

You will complete **12** items in total. Each item shows a image-message pair and the same set of questions.

1. **Answer the Q1 question (Low / High)** about overall persuasiveness for the **image**.
2. **Read the message and look at the image.** Then rate how much **each factor influenced** your overall impression of persuasiveness (1–5 scale).
    - **Sometimes you will see a short-answer box.** If shown, write **1–2 short, specific sentences** explaining whether and how the factor affected persuasiveness.
3. **Answer the Q3 question (Low / High)** about overall persuasiveness for the **message–image pair**.

**There are 2 trap pairs out of 12 items in the study. If you fail to answer them correctly, or if your short-answer responses are deemed insufficient upon overall review, you may not receive the reward.**

### Rating scale

Use a 1–5 scale for each factor:

- **1** = Not at all influential
- **2** = Slightly influential
- **3** = Moderately influential
- **4** = Very influential
- **5** = Extremely influential

### Visual Persuasive Factors definitions

We have identified 9 key factors that can influence visual persuasion, categorized into three levels: perceptual, compositional, and semantic. Please rate **how much each factor influenced** your overall impression of persuasiveness for the given image–message pair.

#### Perceptual-level

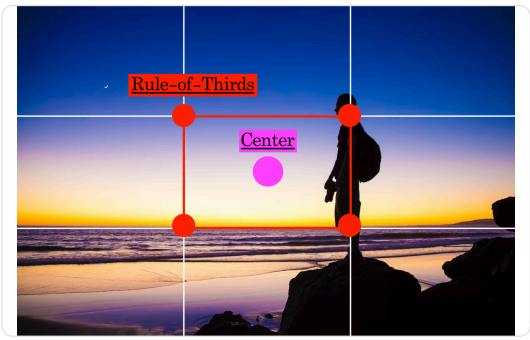
At the perceptual level, you evaluate how factors such as **color and brightness**—which are most immediately apparent—influence persuasiveness.

- **Color** : overall color palette (e.g. dominant hues, saturation, and color contrast).
- **Brightness** : overall lightness

#### Compositional-level

**Center** refers to the middle area of the image. **Rule-of-Thirds (RoT)** refers to the 3×3 grid guideline: subjects placed near the grid intersection points often feel visually emphasized (please refer to the below example).

Figure 10: Instruction page 1 of the human survey interface.




At the compositional level, there are 4 factors:

- **Saliency near the center** : your visual attention toward the center, which influenced how persuasive it felt overall
- **Saliency near RoT points** : your visual attention toward the RoT points, which influenced how persuasive it felt overall
- **Message-relevant region near the center** : the region in image was located near the center, which influenced how persuasive it felt overall
- **Message-relevant region near RoT points** : the region in image was located near the RoT points, which influenced how persuasive it felt overall

---

### Semantic-level

At the semantic level, you evaluate how the presence of various objects in an image influences persuasion. These include **key objects related to the message, humans, and text**.

- **Key object presence** : existence of a key object that represents the message, this could refer to nouns that appear in the message or objects that come to mind when reading the message.
- **Human presence** : people present or not
- **Text presence** : text in the image that contributes to persuasiveness (e.g., slogan, call-to-action, impactful wording)

Example: This image is intended to convey the message, **"Always turn off the stove after use in the kitchen."** The key objects here can be considered to be the 'stove' and 'kitchen'. Additionally, the presence of a human and the 'turn off' text may influence your assessment of the image's persuasiveness.

*Note:* Each high-level factor may be **present or absent** in an image. We will **not explicitly** mark these factors for you. **Please use your own judgment and rate influence (not just presence).**

### Start

**Important: Please enter your Worker ID exactly as shown on MTurk.**
**If you enter the wrong ID, you may not be able to receive reward.**

Enter your Worker ID

Example format: `ABCDEFG123456`

Start

Figure 11: Instruction page 2 of the human survey interface.

# Image-Message Pair

Progress: **1** / **12**

**Image and Message**

**Step 1:** Please answer **Q1** first.
After selecting **High** or **Low**, the message will appear and **Q2** will be unlocked.

## Q1. Visual Persuasion Question

Assuming you do not engage in a specific behavior, please categorize the extent to which you feel prompted to perform that behavior after viewing **only image** as either Low or High.
(After you answer Q1, **the message** will appear and you can proceed to Q2.)

⚠️ **Once you click Low or High, you CANNOT edit this Q1 question.**
**Please review your answers before continuing.**

○ Low ○ High

Figure 12: Step 1 of the human survey interface. Annotators first provide an image-only binary persuasiveness judgment before the message and factor-rating questions are revealed.

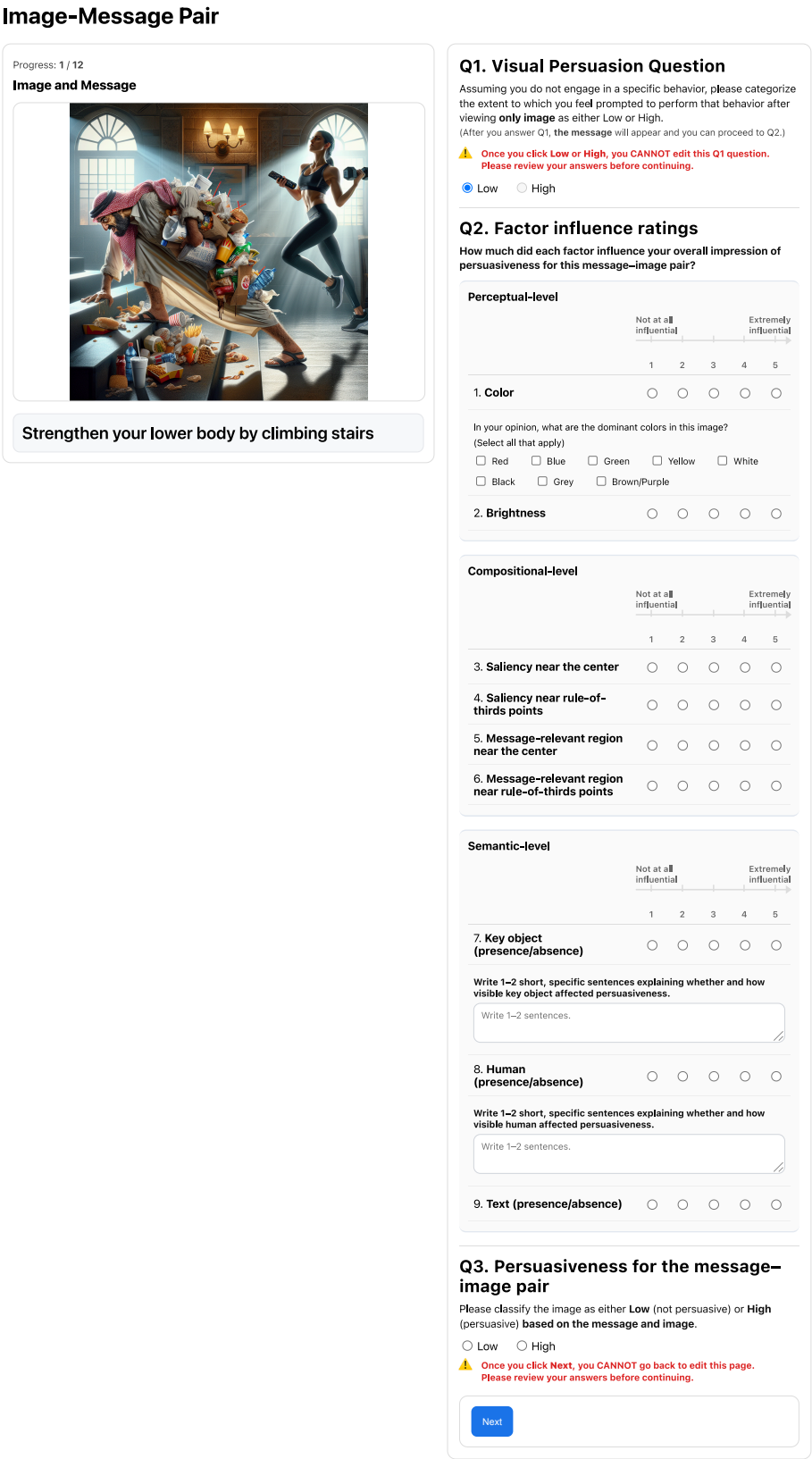



Figure 13: Step 2 of the human survey interface. After answering the image-only question, annotators rate factor-level influence, provide short explanations when applicable, and then judge persuasiveness for the full image-message pair.

# D Experiment Details

All experiments were conducted on a single NVIDIA A6000 GPU. For computational efficiency, all models were loaded using 4-bit quantization, and most experimental single runs (*i.e.,* one seed) completed within four hours. For inference, we employed the default generation configurations provided by each model's respective library; for the GPT models, no additional generation settings were specified beyond the standard API defaults. Detailed model versions and licenses are provided in Tab. 18.

## D.1 Binary Judgment

We use the following baseline prompt for the binary persuasiveness decision. The model receives one image (via the VLM input) and a text message `{message}`. It must output `[[yes]]` if the image is persuasive for the given message, or `[[no]]` otherwise—no other text or reasoning is allowed.

Prompt 1: Baseline (binary judgment).

```
System Instruction:
"You are an AI assistant with expertise in
    analyzing and evaluating the
    persuasiveness of images and messages
    based on general principles of
    communication and psychology."

User Prompt:
"""
Below is an instruction that describes a
    task, paired with an input that provides
    further context. Write a response that
    appropriately completes the request.

### Instruction:
You will perform a task where you predict how
    persuasive certain individuals will find
    an image created from a message. Please
    classify the image as either persuasive
    or not persuasive based on the message
    and image. Respond with a single word:
    'yes' if persuasive, 'no' if not
    persuasive.

### Input:
Message: {message}

please directly output your answer by
    strictly following this format:
    [[answer]], for example: [[yes]]

### Response:
"""
```

## D.2 Knowledge Injection and Reasoning Scaffold

Prompt 2: Factor-aware ognitive injection prompt. Factor instruction details are in the Tab. 19 and Tab. 20.

```
System Instruction: You are an AI assistant
    evaluating whether an image is persuasive
    for the given message.
    "{factor_instruction}"

User Prompt:
"""
Below is an instruction that describes a
    task, paired with an input that provides
    further context. Write a response that
    appropriately completes the request.

### Instruction:
You will perform a task where you predict how
    persuasive certain individuals will find
    an image created from a message. Please
    classify the image as either persuasive
    or not persuasive based on the message
    and image. Respond with a single word:
    'yes' if persuasive, 'no' if not
    persuasive.

### Input:
Message: {message}

please directly output your answer by
    strictly following this format:
    [[answer]], for example: [[yes]]

### Response:
"""
```

| Model | Version | License | Reasoning Effort |
|---|---|---|---|
| GPT-5 (OpenAI, 2025) | `gpt-5-2025-08-07` | Proprietary | minimal (medium for KCC, low for rationale step parsing) |
| GPT-5-mini | `gpt-5-mini-2025-08-07` | Proprietary | minimal (medium for KCC) |
| Gemma-3 (Team et al., 2025) | `gemma-3-27b-it` | Gemma Terms of Use | - |
| Qwen2.5-VL (Bai et al., 2025b) | `Qwen2.5-VL-7B-Instruct` | Apache 2.0 | - |
| LLaVA-1.5 (Liu et al., 2023) | `llava-1.5-13b-hf` | Llama 2 Community License | - |
| LLaVA-Next (Liu et al., 2024) | `llava-v1.6-mistral-7b-hf` | Apache 2.0 | - |

Table 18: Baseline models used for binary persuasiveness judgment.

| Condition | System instruction |
|---|---|
| Color | When making your judgment, explicitly consider the image's overall color properties, including dominant hues, saturation, and overall colorfulness. Use color only as supporting evidence when it plausibly affects persuasion for this message, for example by reinforcing or undermining the message's intended tone, credibility, urgency, or emotional framing. |
| Brightness | When making your judgment, explicitly consider the image's overall brightness and contrast, including how light or dark it appears and how clearly key elements are visible. Use brightness only as supporting evidence when it plausibly affects persuasion for this message. |
| Saliency-Center | In addition to the image, you may be provided with a saliency cue indicating whether visual attention is concentrated near the image center. Saliency refers to where viewers are naturally drawn to look first or most. Treat this as a weak compositional cue, not a decisive rule. |
| Saliency-RoT | In addition to the image, you may be provided with a saliency cue indicating whether visual attention is concentrated near rule-of-thirds points, approximately at the 1/3 and 2/3 intersections of image width and height. Saliency refers to where viewers are naturally drawn to look first or most. Treat this as a weak compositional cue, not a decisive rule. |
| MRR-Center | In addition to the image, identify the message-relevant region, if any, that best corresponds to the message content. Then assess whether this region is located near the image center. The message-relevant region may be absent, ambiguous, or not visually present. If you cannot confidently identify such a region, ignore this consideration and judge normally. |
| MRR-RoT | In addition to the image, identify the message-relevant region, if any, that best corresponds to the message content. Then assess whether this region is located near rule-of-thirds points, approximately at the 1/3 and 2/3 intersections of image width and height. The message-relevant region may be absent, ambiguous, or not visually present. If you cannot confidently identify such a region, ignore this consideration and judge normally. |
| Key Object | Pay strong attention to whether the image depicts the message's key object(s); then judge whether and how much the presence, clarity, and prominence of those key object(s) should influence persuasiveness in this specific case. If there are no visible key object(s), proceed without using this cue. |
| Human | When making your judgment, consider whether people or humans are visually present, including faces, bodies, or crowds, but treat human presence as neutral by default. Do not assume that including people automatically increases persuasiveness. Only use human presence as supporting evidence if it clearly contributes to persuasion for this specific message, for example through social proof, demonstration, emotion, credibility, or context. |
| Text | When making your judgment, explicitly check whether the image contains visible text, such as captions, slogans, labels, overlays, signs, or dense paragraphs. Use text only as supporting evidence when it plausibly affects persuasion for this message. If there is no visible text, proceed without using this cue. |

Table 19: Factor-specific system instructions used for factor-aware cognitive injection.

| Condition | System instruction |
|---|---|
| All Factors | You are an AI assistant evaluating whether an image is persuasive for the given message.<br>When making your judgment, jointly consider the following nine visual persuasion factors. Use them as supporting evidence, not as rigid rules. No single factor should dominate unless it clearly changes how persuasive the image is for the specific message.<br>1. Color: Consider the image's overall color properties, including dominant hues, saturation, and overall colorfulness. Use color as evidence when it reinforces or undermines the message's tone, credibility, urgency, or emotional framing.<br>2. Brightness: Consider the image's overall brightness and contrast, including whether key elements are clearly visible. Use brightness as evidence when lightness, darkness, or visibility plausibly affects persuasion.<br>3. Message-Relevant Region Centering: Identify the message-relevant visual region, if any, that best corresponds to the message content. Assess whether this region is located near the image center. If the relevant region is absent or ambiguous, ignore this cue.<br>4. Message-Relevant Region Rule of Thirds: Identify the message-relevant visual region, if any, and assess whether it is located near rule-of-thirds points, approximately at the 1/3 and 2/3 intersections of image width and height. If the relevant region is absent or ambiguous, ignore this cue.<br>5. Saliency Centering: Consider whether visual attention is concentrated near the image center. Treat center saliency as a weak compositional cue, not a decisive rule.<br>6. Saliency Rule of Thirds: Consider whether visual attention is concentrated near rule-of-thirds points. Treat rule-of-thirds saliency as a weak compositional cue, not a decisive rule.<br>7. Key Object Presence: Pay strong attention to whether the image clearly depicts the key object(s) implied by the message. Clear depiction of the core object can increase persuasiveness, while absence or ambiguity can weaken it.<br>8. Human Presence: Consider whether people or humans are visually present, including faces, bodies, or crowds. Treat human presence as neutral by default. Use it as evidence only when it contributes to persuasion through social proof, demonstration, emotion, credibility, or context.<br>9. Text Presence: Check whether the image contains visible text such as captions, slogans, labels, overlays, signs, or dense paragraphs. Use text as evidence when it helps clarify or support the message.<br>If a cue is not visible, absent, or ambiguous, simply proceed without using that cue. After considering all applicable factors together, decide whether the image is persuasive for the message. |

Table 20: System instruction used for the All Factors condition.

| Model | Condition | Acc. | Prec. (Δ) | Rec. (Δ) | F1 (Δ) |
|---|---|---|---|---|---|
| **Panel A. Baseline** | | | | | |
| GPT-5 | Baseline | 69.57 | 51.01 (0.00) | 98.88 (0.00) | 67.30 (0.00) |
| GPT-5-mini | Baseline | 60.14 | 44.13 (0.00) | 97.19 (0.00) | 60.70 (0.00) |
| Gemma3 | Baseline | 63.88 | 46.33 (0.00) | 88.76 (0.00) | 60.89 (0.00) |
| Qwen2.5-VL | Baseline | 54.98 | 40.51 (0.00) | 89.89 (0.00) | 55.85 (0.00) |
| **Panel B. Factor-wise Cognitive Injection** | | | | | |
| *GPT-5* | | | | | |
| GPT-5 | CI(Color) | 67.08 | 49.03 ↓1.98 | 98.88 (0.00) | 65.55 ↓1.75 |
| GPT-5 | CI(Brightness) | 67.44 | 49.30 ↓1.71 | 99.44 ↑0.56 | 65.92 ↓1.38 |
| GPT-5 | CI(Sal-Center) | 68.51 | 50.14 ↓0.87 | 98.88 (0.00) | 66.54 ↓0.76 |
| GPT-5 | CI(Sal-RoT) | 70.11 | 51.46 ↑0.45 | 98.88 (0.00) | 67.69 ↑0.39 |
| GPT-5 | CI(Reg-Center) | 68.86 | 50.43 ↓0.58 | 98.31 ↓0.57 | 66.67 ↓0.63 |
| GPT-5 | CI(Reg-RoT) | 69.22 | 50.72 ↓0.29 | 98.88 (0.00) | 67.05 ↓0.25 |
| GPT-5 | CI(KeyObj) | 70.11 | 51.47 ↑0.46 | 98.31 ↓0.57 | 67.57 ↑0.27 |
| GPT-5 | CI(Human) | 72.60 | 53.70 ↑2.69 | 97.75 ↓1.13 | 69.32 ↑2.02 |
| GPT-5 | CI(Text) | 65.30 | 47.72 ↓3.29 | 100.00 ↑1.12 | 64.61 ↓2.69 |
| *GPT-5-mini* | | | | | |
| GPT-5-mini | CI(Color) | 61.92 | 45.29 ↑1.16 | 97.19 (0.00) | 61.79 ↑1.09 |
| GPT-5-mini | CI(Brightness) | 64.06 | 46.79 ↑2.66 | 98.31 ↑1.12 | 63.41 ↑2.71 |
| GPT-5-mini | CI(Sal-Center) | 62.10 | 45.33 ↑1.20 | 95.51 ↓1.68 | 61.48 ↑0.78 |
| GPT-5-mini | CI(Sal-RoT) | 63.17 | 46.07 ↑1.94 | 95.51 ↓1.68 | 62.16 ↑1.46 |
| GPT-5-mini | CI(Reg-Center) | 61.03 | 44.70 ↑0.57 | 97.19 (0.00) | 61.24 ↑0.54 |
| GPT-5-mini | CI(Reg-RoT) | 61.39 | 44.91 ↑0.78 | 96.63 ↓0.56 | 61.32 ↑0.62 |
| GPT-5-mini | CI(KeyObj) | 61.03 | 44.70 ↑0.57 | 97.19 (0.00) | 61.24 ↑0.54 |
| GPT-5-mini | CI(Human) | 61.74 | 45.24 ↑1.11 | 98.88 ↑1.69 | 62.08 ↑1.38 |
| GPT-5-mini | CI(Text) | 62.28 | 45.33 ↑1.20 | 97.19 (0.00) | 62.01 ↑1.31 |
| *Gemma3* | | | | | |
| Gemma3 | CI(Color) | 71.00 | 52.63 ↑6.30 | 84.27 ↓4.49 | 64.79 ↑3.90 |
| Gemma3 | CI(Brightness) | 68.15 | 49.84 ↑3.51 | 87.64 ↓1.12 | 63.54 ↑2.65 |
| Gemma3 | CI(Sal-Center) | 68.86 | 50.50 ↑4.17 | 85.96 ↓2.80 | 63.62 ↑2.73 |
| Gemma3 | CI(Sal-RoT) | 70.28 | 51.85 ↑5.52 | 86.52 ↓2.24 | 64.84 ↑3.95 |
| Gemma3 | CI(Reg-Center) | 65.84 | 47.88 ↑1.55 | 88.76 (0.00) | 62.20 ↑1.31 |
| Gemma3 | CI(Reg-RoT) | 66.01 | 48.07 ↑1.74 | 91.01 ↑2.25 | 62.91 ↑2.02 |
| Gemma3 | CI(KeyObj) | 67.08 | 48.85 ↑2.52 | 83.71 ↓5.05 | 61.70 ↑0.81 |
| Gemma3 | CI(Human) | 68.86 | 50.47 ↑4.14 | 89.89 ↑1.13 | 64.65 ↑3.76 |
| Gemma3 | CI(Text) | 67.79 | 49.54 ↑3.21 | 89.89 ↑1.13 | 63.87 ↑2.98 |
| *Qwen2.5-VL* | | | | | |
| Qwen2.5-VL | CI(Color) | 62.81 | 45.35 ↑4.84 | 84.83 ↓5.06 | 59.10 ↑3.25 |
| Qwen2.5-VL | CI(Brightness) | 61.74 | 44.64 ↑4.13 | 86.52 ↓3.37 | 58.89 ↑3.04 |
| Qwen2.5-VL | CI(Sal-Center) | 61.57 | 44.41 ↑3.90 | 84.83 ↓5.06 | 58.30 ↑2.45 |
| Qwen2.5-VL | CI(Sal-RoT) | 63.17 | 45.62 ↑5.11 | 84.83 ↓5.06 | 59.33 ↑3.48 |
| Qwen2.5-VL | CI(Reg-Center) | 65.48 | 47.47 ↑6.96 | 84.27 ↓5.62 | 60.73 ↑4.88 |
| Qwen2.5-VL | CI(Reg-RoT) | 63.88 | 46.06 ↑5.55 | 82.02 ↓7.87 | 58.99 ↑3.14 |
| Qwen2.5-VL | CI(KeyObj) | 58.90 | 42.82 ↑2.31 | 88.76 ↓1.13 | 57.77 ↑1.92 |
| Qwen2.5-VL | CI(Human) | 65.48 | 47.01 ↑6.50 | 70.79 ↓19.10 | 56.50 ↑0.65 |
| Qwen2.5-VL | CI(Text) | 65.84 | 47.71 ↑7.20 | 82.02 ↓7.87 | 60.33 ↑4.48 |
| **Panel C. All-factor Cognitive Injection** | | | | | |
| GPT-5 | CI(All) | 69.93 | 51.35 ↑0.34 | 96.07 ↓2.81 | 66.93 ↓0.37 |
| GPT-5-mini | CI(All) | 60.50 | 44.33 ↑0.20 | 96.63 ↓0.56 | 60.78 ↑0.08 |
| Gemma3 | CI(All) | 58.57 | 43.31 ↓3.02 | 94.94 ↑6.18 | 59.30 ↓1.59 |
| Qwen2.5-VL | CI(All) | 62.10 | 44.87 ↑4.36 | 85.96 ↓3.93 | 58.96 ↑3.11 |

Table 21: Results of factor-wise and all-factor cognitive injection on VLM persuasive judgment performance. *Notes.* Δ next to Precision, Recall, and F1 indicates the absolute percentage-point change from each model's own baseline Precision, Recall, and F1 scores, respectively.

Prompt 3: Knowledge-conditioned chain.

```
System Instruction:
"You are an AI assistant with expertise in
    analyzing and evaluating the
    persuasiveness of images and messages
    based on general principles of
    communication and psychology."

User Prompt:
"""
Below is an instruction that describes a
    task, paired with an input that provides
    further context. Write a response that
    appropriately completes the request.

### Instruction:
Decide whether an image (if present or
    implied) created from the message would
    be persuasive to likely viewers.

Before reasoning, note:
Key objects are supporting cues, not
    determining factors.
- They can provide additional hints for
    interpreting the message but are not
    required for persuasiveness.
- Their presence may support or reinforce the
    message when contextually aligned, but
    does not guarantee higher persuasiveness.
- Their absence does not necessarily reduce
    persuasiveness unless the message
    explicitly depends on them.

Use the following reasoning steps, and show
    your reasoning explicitly before giving
    the final answer:
1) Extract candidate key objects.
2) Analyze visual-message alignment.
3) Make a final judgment on whether the image
    effectively expresses or reinforces the
    message.

### Input:
Message: {message}

### Response:
Reasoning: <your reasoning here>
Answer: [[yes]] or [[no]]
"""
```

Prompt 4: Aligned key object context, only used when a key object exists.

```
System Instruction:
"You are an AI assistant with expertise in
    analyzing and evaluating the
    persuasiveness of images and messages
    based on general principles of
    communication and psychology."

User Prompt:
"""
Below is an instruction that describes a
    task, paired with an input that provides
    further context. Write a response that
    appropriately completes the request.

### Instruction:
You will perform a task where you predict how
    persuasive certain individuals will find
    an image created from a message. Respond
    with a single word: 'yes' if persuasive,
    'no' if not persuasive.
IMPORTANT: The following key object(s) were
    algorithmically detected in the image and
    should be treated as verified evidence of
    presence. Use these objects only as
    evidence for what appears in the image;
    your final decision should still be based
    on overall persuasiveness.

### Input:
Message: {message}
Key Object: {key_object_json}

please directly output your answer by
    strictly following this format:
    [[answer]], for example: [[yes]]

### Response:
"""
```

## D.3 Rationale Generator

We use a rationale-generation prompt, where the model generates rationales without access to the human persuasiveness label. This setting allows us to examine whether models can explain the persuasive role of detected key objects without being conditioned on the ground-truth judgment.

Prompt 5: Key object rationale generation.

```
System Instruction:
"You are an expert visual persuasion analyst.
    Given an image, a message, and detected
    key object tokens, explain concisely how
    the object(s) likely affect
    persuasiveness, focusing on message-image
    alignment, object salience/clarity,
    emotional/moral cues, and credibility."

User Prompt:
"""
Analyze the following image-message pair.

### Instructions:
1) Provide a concise overall rationale for
    how persuasive (or not) the pair is
    likely to be.
2) For EACH key object in the list, write 1-2
    short, specific sentences explaining
    whether and how that object increased or
    decreased persuasiveness.
3) Be concrete (what is visible + how it
    supports/undermines the message). Avoid
    speculation beyond the image.
4) OUTPUT **ONLY** a JSON object with this
    exact schema (no extra text):
{{
  "per_object": {{"<object1>": "<1-2 short
    sentences>", "<object2>": "<...>"}},
}}

### Inputs:
message: {message}
key_objects: {key_objects}
```

```
Return JSON only.

### Response:
"""
```

Prompt 6: Key object rationale, only used when a key object exists. The {key_objects} field contains the per-object rationales.

```
System Instruction:
"You are an AI assistant with expertise in
    analyzing and evaluating the
    persuasiveness of images and messages
    based on general principles of
    communication and psychology."

User Prompt:
"""
Below is an instruction that describes a
    task, paired with an input that provides
    further context. Write a response that
    appropriately completes the request.

### Instruction:
You will perform a task where you predict how
    persuasive certain individuals will find
    an image created from a message. You are
    also given a verbatim JSON map of
    per-object rationales extracted from
    prior analysis. Use these rationales
    as-is as additional context. Please
    classify the image as either persuasive
    or not persuasive based on the message
    and image. Respond with a single word:
    'yes' if persuasive, 'no' if not
    persuasive.

### Input:
Message: {message}
Key Objects: {key_objects}  # for reference

please directly output your answer by
    strictly following this format:
    [[answer]], for example: [[yes]]

### Response:
"""
```

### D.4 Step-Wise Rationale Parsing Details

**Parsing Procedure Details.** For the step-wise rationale analysis, we used GPT-5 to parse each generated key object rationale into three functional spans. We used reasoning effort set to `low`, rather than `minimal`, because we empirically found that `low` produced more reliable and stable span annotations while still preserving the original rationale text. The parser was instructed to insert only step tags into the original rationale, without paraphrasing, deleting, or reordering any part of the text. The exact prompt used for parsing is shown in Prompt 7.

Prompt 7: Step-wise rationale parsing.

```
System Instruction:
"You are a strict text editor.
Your job is to insert ONLY the tags
    <S1>...</S1>, <S2>...</S2>, <S3>...</S3>
    into the given text.
You MUST NOT change, paraphrase, reorder,
    delete, or add any characters of the
    original text.
The only allowed edits are inserting these
    tags at character boundaries.

You must segment the text into exactly three
    contiguous spans:
- <S1> Factor Detection / Evidence (what is
    observed or stated as evidence)
- <S2> Evidence-to-Claim Mapping (how the
    evidence supports or conflicts with the
    message/claim)
- <S3> Persuasive Judgement / Deduction
    (final evaluative conclusion)

Segmentation policy (IMPORTANT):
- You may use these tags MULTIPLE TIMES.
- The order does NOT have to be S1 -> S2 ->
    S3. It can be in ANY ORDER depending on
    the text structure.

Hard constraints:
1) All original text must be covered by
    exactly one of S1/S2/S3. No text outside
    the tags.
2) Do NOT insert any other tokens, headings,
    bullet points, or explanations.
3) Output must be valid JSON only."

User Prompt:
"""
obj_key: {obj_key}

rationale_text (DO NOT EDIT ANY CHARACTERS,
    ONLY INSERT TAGS):
{rationale}

Return JSON only in this schema:
{
  "obj_key": "{obj_key}",
  "tagged_rationale": "<TAGGED_RATIONALE>"
}
"""
```

**Parsing Examples.** Fig. 14 shows representative parsed rationales across models. The examples illustrate that rationale functions are not always separated by sentence boundaries: a single sentence can contain both evidence and mapping, while longer rationales may alternate between evidence and interpretation before reaching the final judgment.

**False Positive Analysis.** We further analyze step-wise ablation from the perspective of false positives. As shown in Fig. 16, removing rationale components can reduce the tendency of models to incorrectly classify GT-Low items as persuasive, but the effect is not uniform across steps or models. In particular, ablating the evidence-to-claim mapping step, either alone or together with evidence detec-

S1: Evidence S2: Mapping S3: Judgement

**GPT5 | keyobj=military**
**Message: Participate in military events**

A large, diverse formation of uniformed service members beneath a massive U.S. flag with a jet flyover makes the military theme unmistakable and ceremonial. Clear uniforms, badges, and formation boost credibility and patriotic emotion that encourages participation, though the idealized staging may feel propagandistic to some.

**GPT5 | keyobj=car**
**Message: Watch out for bicycles and motorcycles when getting out of your car**

The car is central and front-facing, with its door visibly striking the cyclist, concretely illustrating the exact hazard of opening a door without checking. The simple cartoon rendering makes the point clear but slightly reduces realism and urgency.

**GPT5-mini | keyobj=homeless**
**Message: Donate to homeless support programs**

The person sitting on the sidewalk with a cardboard sign reading 'HOMELESS' makes the need explicit and increases persuasiveness by invoking empathy and providing a clear donation target. Nearby items — a food tray and a cup — visually reinforce scarcity and immediacy of need, adding credibility to the appeal.

**GPT5-mini | keyobj=legs**
**Message: Do not shake your legs while studying**

The legs are shown both stationary and as a blurred duplicate, making leg movement highly visible and easy to interpret. This clear contrast ties leg shaking directly to motion/distraction in a study setting, which supports the 'do not shake your legs while studying' message.

**Gemma3 | keyobj=military**
**Message: Participate in military events**

The highly visible and diverse group of military personnel in uniform lends credibility to the message, implying broad support and participation. The inclusion of families and a welcoming, sunlit scene evokes positive emotional associations with military service, encouraging participation.

**Gemma3 | keyobj=photos**
**Message: Do not upload your photos on social media**

The image prominently features a photo icon *within* a folder icon, visually representing photos as contained and potentially private. This supports the message by suggesting a contrast to the public nature of social media, subtly implying photos are safer kept 'contained'.

**Qwen2.5-VL | keyobj=airport**
**Message: Declare agricultural products at the airport when bringing them from abroad**

The image of a busy airport terminal with people waiting in line at customs counters strongly aligns with the message about declaring agricultural products when bringing them from abroad. The presence of multiple travelers and the clear signage for customs emphasizes the importance of adhering to customs regulations.

**Qwen2.5-VL | keyobj=clothes**
**Message: Dry your clothes naturally**

The clothes hanging on lines suggest a natural drying process, which aligns with the message of 'Dry your clothes naturally'. However, the presence of industrial smokestacks in the background creates a stark contrast, undermining the natural aspect of the message.

Figure 14: Examples of step-wise rationale parsing. Each rationale is shown with text-preserving span annotations for **S1** evidence detection, **S2** evidence-to-claim mapping, and **S3** persuasive judgment.

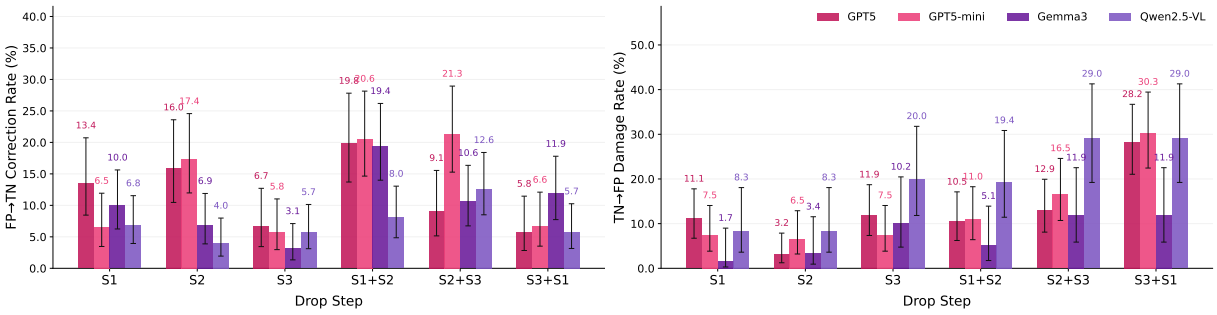



Figure 15: Correction–damage trade-off, comparing FP-to-TN corrections against TN-to-FP errors.

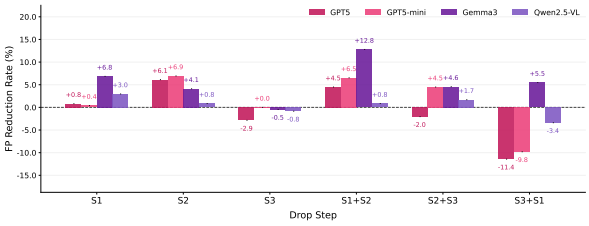



Figure 16: Reduction in false positive rate after removing each rationale step.

tion, often yields the largest false-positive reduction, suggesting that models may overestimate persuasiveness when their rationales construct overly strong links between visual evidence and the target message. Fig. 15 further shows that this improvement should be interpreted as a correction–damage trade-off: some ablations successfully convert false positives into true negatives, while others also introduce new TN-to-FP errors. Overall, these results indicate that false positives are caused by intermediate reasoning steps that over-align visual cues with the persuasive claim.