Do Vision–Language Models Understand Visual Persuasiveness?

Gyuwon Park
Department of Computer Science and Engineering, UNIST
gyuwon12@unist.ac.kr

Abstract

Recent advances in vision–language models (VLMs) have enabled impressive multi-modal reasoning and understanding. Yet, whether these models truly grasp *visual persuasion*—how visual cues shape human attitudes and decisions—remains unclear. To probe this question, we construct a high-consensus dataset for binary persuasiveness judgment and introduce the taxonomy of *Visual Persuasive Factors (VPFs)*, encompassing low-level perceptual, mid-level compositional, and high-level semantic cues. We also explore cognitive steering and knowledge injection strategies for persuasion-relevant reasoning. Empirical analysis across VLMs reveals a recall-oriented bias—models over-predict high persuasiveness—and weak discriminative power for low/mid-level features. In contrast, high-level semantic alignment between message and object presence emerges as the strongest predictor of human judgment. Among intervention strategies, simple instruction or unguided reasoning scaffolds yield marginal or negative effects, whereas concise, object-grounded rationales significantly improve precision and F1 scores. These results indicate that VLMs core limitation lies not in recognizing persuasive objects but in linking them to communicative intent.

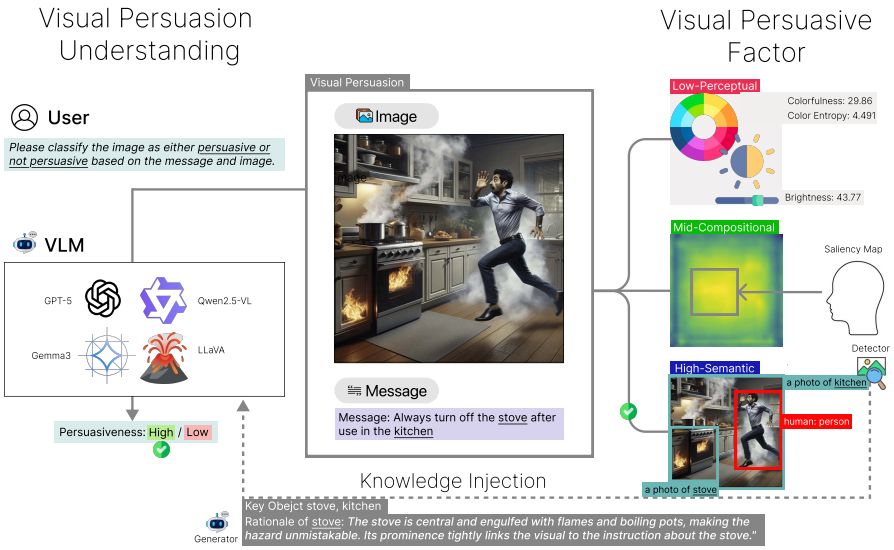


Figure 1: Overview of framework for **Visual Persuasion Understanding** (left) and the proposed **Visual Persuasive Factors** (right). The model (VLM) receives an image–message pair and judges persuasiveness, while human-identified factors—spanning low-perceptual, mid-compositional, high-semantic level, and high-semantic cues—enable deeper analysis of what drives persuasive reasoning.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: VLM4RWD.

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 [1]. Visuals are not just decorative—they reliably shape attitudes and behavior, as classic persuasion theory [2, 3] and public-health evidence on pictorial warnings have shown. In real-world deployment, a VLM that can *judge* and *explain* visual persuasiveness can (i) help select or A/B-triage more effective creatives under latency/cost constraints and (ii) act as a guardrail by flagging potentially manipulative imagery, a critical need as images systematically bias truth judgments and fuel visual misinformation [4, 5].

In the AI domain, research on visual persuasion has taken two primary directions: understanding how persuasive an image is and generating images that effectively convey persuasive intent. Early efforts in *visual persuasive understanding*—such as [6]—analyzed handcrafted visual attributes to infer social and psychological intent. Recent datasets and models, including the *Personalized Visual Persuasion* (PVP) dataset, extend this line of work by pairing 28,454 persuasive images with 596 messages annotated for nine persuasion strategies [1]. However, existing work provides limited insight into how well VLMs actually understand visual persuasiveness—*why* certain images are persuasive or which persuasive factors contribute to their effects remains unclear.

In this work, we investigate whether state-of-the-art VLMs possess a meaningful notion of *visual persuasive understanding*. While visual persuasion is inherently subjective—its effects often arising in ambiguous, low-agreement settings shaped by individual personality—such factors (e.g., psychological profiles) are difficult to ascertain in advance. This motivates us to explore whether robust persuasive factors can be established from the message or visual cues instead. We therefore focus on cases with *high human agreement*, where the persuasive intent is clear and universally recognizable. However, even in these seemingly straightforward scenarios, we find that VLMs struggle to exhibit genuine *persuasiveness understanding*: they tend to make indiscriminately positive judgments, revealing a recall-oriented bias. We conceptualize this as the ability to identify the visual factors that drive persuasiveness and to integrate them into coherent judgments.

Building on this formulation, we propose a taxonomy of *Visual Persuasive Factors* (VPFs) grounded in cognitive psychology, spanning low-level perceptual features, mid-level compositional features, and high-level semantic features. Our contributions are threefold: (1) we perform a systematic evaluation of VLMs on visual persuasion judgment and identify a pervasive recall-oriented bias; (2) we define and extract VPFs and use them to analyze human and model judgments, revealing which visual cues are most predictive of persuasiveness; and (3) we explore knowledge-injection strategies that encourage models to attend to visual cues.

2 Related Work

2.1 Persuasion with AI

Research on computational persuasion began with text-only settings. Early dialogue corpora such as *Persuasion for Good* enabled systems that generate persuasive responses with hand-crafted strategies [7]; follow-up work incorporated affect and user traits to produce more empathetic appeals and to probe whether LLM-generated messages are judged as persuasive by humans [8–10].

Visual persuasion entered the AI literature with early computer-vision formulations that linked image content to communicative intent [1, 11]. With the rise of multi-modal LMs, large-scale datasets now couple images with persuasive messages; notably, the *Personalized Visual Persuasion* (PVP) [1] corpus frames prediction as scoring the persuasiveness of image–text pairs. While such work establishes the task, it typically optimizes a scalar score without exposing *which* visual cues drive judgments.

Recent evaluation protocols move beyond end-to-end scoring to measure persuasiveness directly, for instance via human-in-the-loop or controlled prompting setups that quantify model persuasion or its safety implications [12, 13]. Our study follows this direction in the visual domain: we start from binary judgments for robustness, then analyze *why* models respond as they do by decomposing images into low-, mid-, and high-level *visual persuasive factors* and comparing these factors against human decisions.

2.2 Visual Persuasion Factor in Cognitive Psychology

Our factor taxonomy draws on findings from cognitive psychology and communication theory. The Elaboration Likelihood Model (ELM) [2] of persuasion posits two distinct routes: a *central* route in which audiences carefully elaborate on a message's content, and a *peripheral* route in which superficial cues guide judgments. Our visual persuasion typically operates on the peripheral route, leveraging affective and intuitive signals such as color, contrast and composition.

Numerous behavioral studies demonstrate that low-level visual properties modulate cognition; for example, *color* is widely recognized as a fundamental component of human visual perception and persuasion [14–17]. In advertisements and campaigns, which are the main domains of visual persuasion, certain colors also act as a strong persuasion factor [18–20].

Human visual perception (eye-tracking) is crucial in visual persuasion and is often modeled using saliency prediction [21–23]. Beyond simple features like color, composition and visual hierarchy also guide attention, as exemplified by the well-known center bias in natural scenes [24, 25]. Saliency prediction can thus be used to analyze how such compositional principles guide a viewer's scan path and ultimately contribute to persuasion.

At the high level, persuasiveness depends on whether the objects in the image align semantically with the intended message and whether the overall narrative is coherent; recognizing such relations requires the ability to select the relevant objects and connect them to an argument structure [26]. The core object may include a product, text, or person [27–29].

2.3 High-Level Vision–Language Reasoning

Recent foundation models [30–32] have rapidly evolved beyond language-only applications, unlocking impressive performance on complex, high-dimensional tasks. Also, work on chain-of-thought prompting demonstrates that allowing a model to generate intermediate reasoning steps significantly improves its ability to solve arithmetic, commonsense and symbolic problems [33–38].

Vision–language models (VLMs) extend this paradigm by coupling LLMs with visual encoders [39–42], allowing multi-modal inputs. Recent advances in VLMs have moved beyond literal object recognition toward interpreting abstract and symbolic meanings. For example, models are now evaluated on understanding *visual metaphors*—such as a dove representing peace or a storm symbolizing conflict—which require reasoning that bridges perception and semantics [43, 44]. Moreover, in multi-modal reasoning, research has explored integrating reasoning chains into image generation pipelines, applying the CoT technique to visual tasks. [45, 46]

3 From VLMs Persuasive Bias to Visual Persuasive Factors

We transition from aggregate accuracy to *why* models decide as they do. We first document a systematic bias, then introduce a factorized rubric that explains it.

3.1 Identifying VLM's Persuasiveness

To assess whether VLMs understand visual persuasion, we extend the PVP dataset [1] by filtering messages and framing persuasion as a binary judgment task (image + message $\rightarrow$ {*high*, *low*}; overview in Figure 1 and details in the Appendix A). We report accuracy, precision, recall, and F1 score in Table 2 baseline column. Strong VLMs [31, 41, 42] consistently show a recall-oriented bias: they predict *high* too often—near-perfect recall with inflated false positives—while weaker open models (e.g., LLaVA variants) underperform across the board. We hypothesize that these low performance and bias stem from *misinterpretation of visual cues* in the image–text pair: models key on signals correlated with persuasiveness but not causally diagnostic. This is because the poor performance, even on a high-agreement dataset, is attributed to the difficulty in linking cues from the message or the image itself to persuasiveness.

3.2 Defining Visual Persuasive Factors

Our *Visual Persuasive Factor* decomposes the persuasive content of an image into three levels inspired by cognitive psychology.

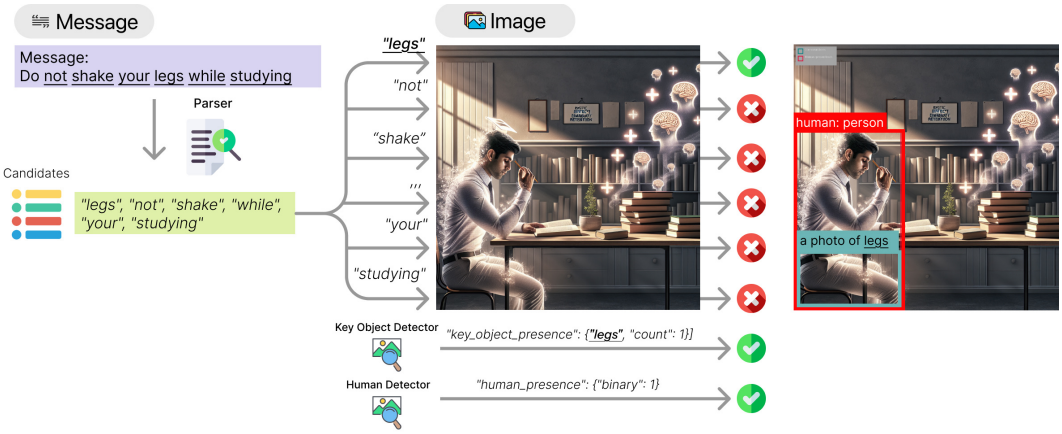


Figure 2: Overview of the high-level feature extraction pipeline. The framework detects whether semantically relevant *key objects* mentioned in the message are visually present in the image and identifies *human presence* through dedicated detectors.

- **Low—Perceptual:** capture immediate, subconscious impressions; we include global color attributes such as colorfulness and entropy and overall brightness.
- **Mid—Compositional:** describe the spatial organization of visual elements, including attention concentration, distribution, center bias and rule-of-thirds alignment, which can be measured from saliency maps.
- **High—Semantic/Contextual:** require recognizing objects and their contextual roles; we assess whether key nouns mentioned in the message are present in the image, reasoning that alignment between textual and visual semantics is crucial for persuasion.

For each image–message pair we extract the factors and build a robust human-annotated dataset under the assumption that annotators largely agree on what makes an image persuasive.

4 Dissecting Visual Persuasion Bias in VLMs

4.1 VPF Feature Extraction

To quantitatively measure low-level perceptual features, we extracted metrics related to image color and brightness. Specifically, we used the colorfulness metric proposed by Hasler-Süsstrunk [47] to capture the color intensity and calculated the color entropy from the CIELAB space to measure the color diversity and utilized the L^* channel (also from CIELAB) to quantify perceptual brightness. These metrics allow us to quantify the perceptual properties of an image's color palette. Detailed definitions and formal equations for these metrics, along with examples of analyzes, are provided in the Appendix B.1.

For mid-level compositional features, we analyzed how spatial arrangement guides human attention using saliency maps extracted from DeepGaze IIE [22]. From these maps, we derived four metrics to characterize visual attention: Top-p Mass Area ($A@p$, $p = 0.85$) to measure attention concentration, Normalized Saliency Entropy (H_{sal}) for attention distribution, Center Bias Index (CBI) to quantify focus on the image center, and Thirds Hotspot Coverage (T3) to assess alignment with the rule of thirds. The specific calculations for each metric and extracted examples can be found in Appendix B.2.

Lastly, we extracted two high-level features related to semantic coherence. The first task, Key Object Presence, identifies whether key concepts mentioned in the accompanying text are visually represented in the image. This involves parsing the text to extract main nouns and verifying their corresponding visual instances. The second task, Human Presence, determines whether one or more people appear in the image. Together, these features capture the semantic alignment between modalities, treating both text-derived concepts and the presence of humans as critical persuasive objects. To evaluate their contribution to perception-level outcomes, we represented key object and human presence as binary indicators and fitted logistic regression models to estimate their effects on

Table 1: Human vs. VLM analyses across three feature levels. FP: human-labeled *low* but predicted *high*.

Group	Row	Low-Level			Mid-Level				High-Level	
		Colorfulness	Color Entropy	Brightness	$A@p$	H_{sal}	CBI	T3	KeyObj (Δ%)	Human (Δ%)
Panel A: Human (ground truth / analysis)										
Human	Y = High	43.99	5.156	45.08	.4453	.9619	.3558	.2100	3.284***[1]	1.222
Human	Y = Low	50.03	4.608	52.66	.4393	.9592	.3289	.1778		
Panel B: VLM predictions										
GPT-5	$\hat{Y}$ = High	47.25	5.005	48.59	.4409	.9607	.3427	.1975	2.868***(↓12.67)	1.562*(↑27.82)
	$\hat{Y}$ = Low	49.49	4.427	52.91	.4416	.9590	.3291	.1729		
GPT-5-mini	$\hat{Y}$ = High	46.55	4.915	48.67	.4416	.9607	.3423	.1958	2.696***(↓17.89)	1.584*(↑29.62)
	$\hat{Y}$ = Low	51.73	4.474	53.92	.4403	.9585	.3263	.1701		
Gemma3	$\hat{Y}$ = High	46.53	4.927	47.83	.4400	.9606	.3463	.1991	2.041***(↓37.86)	1.881**(↑53.93)
	$\hat{Y}$ = Low	50.57	4.557	54.01	.4430	.9591	.3237	.1709		
Qwen2.5-VL	$\hat{Y}$ = High	47.03	4.960	48.63	.4409	.9605	.3467	.1986	2.733***(↓16.77)	2.479***(↑102.86)
	$\hat{Y}$ = Low	50.70	4.361	54.11	.4418	.9588	.3155	.1629		
Panel C: VLMs (by error type; recall-oriented bias surfaced by FP/TP split)										
GPT-5	FP (low→high)	50.82	4.840	52.30	.4357	.9593	.3301	.1848	–	–
	TP (correct)	43.83	5.164	45.03	.4461	.9621	.3555	.2110	–	–
GPT-5-mini	FP	48.85	4.717	51.64	.4384	.9598	.3322	.1847	–	–
	TP	43.65	5.166	44.92	.4459	.9620	.3555	.2110	–	–
Gemma3	FP	49.30	4.766	50.68	.4371	.9596	.3362	.1881	–	–
	TP	43.32	5.114	44.53	.4437	.9618	.3580	.2129	–	–
Qwen2.5-VL	FP	49.32	4.811	51.22	.4385	.9596	.3377	.1885	–	–
	TP	43.66	5.177	44.83	.4450	.9620	.3592	.2143	–	–

Notes. Δ% denotes the proportional change relative to the ground truth baseline. For the high-level Human column, ground-truth OR were not statistically significant; thus, comparisons are interpreted descriptively rather than inferentially. 95% CI and other information are reported in Appendix B.3.

perceived persuasiveness. High-level columns in Table 1 report odds ratios (OR) from these logistic regressions with heteroskedasticity-robust standard errors. A concise overview of this pipeline is provided in Figure 2 and details are provided in Appendix B.3.

4.2 Analysis of Visual Persuasive Factor

Due to significant instruction following failures and poor output stability from the LLaVA variants [39, 40], we omit them from the analysis in Table 1 to ensure valid comparisons. While we report LLaVA results in Table 2 for completeness, our primary interpretation focuses on the other models for these same reasons.

Comparison with Humans (Low-Level). In Table 1, we compare how humans and VLMs assess persuasiveness through low-level perceptual cues. Humans tend to perceive persuasive images as having *slightly lower colorfulness and brightness but higher color entropy*, reflecting weaker overall color intensity per the Hasler–Süsstrunk definition of colorfulness and richer chromatic diversity in color distribution (see example in Appendix B.1). VLMs broadly follow this direction, suggesting a partial alignment in perceptual tendencies, though the differences remain marginal and fall mostly within the "*moderately* to *averagely* colorful" regime. False positives, however, are typically brighter and more colorful with lower entropy. This implies that models are using these simple perceptual properties as a misleading heuristic for persuasion. In fact, even within the human data (Panel A), the numerical differences between 'High' and 'Low' persuasiveness are marginal and overlapping. These weak trends indicate that low-level cues offer contextual background rather than discriminative evidence; we think of them as not the primary factor humans use to judge persuasiveness.

Comparison with Mid-Level Composition. Across models, images labeled *high* versus *low* show only minute differences in global saliency statistics, and FP/TP splits are similarly tight—true positives exhibit at most a slightly stronger center-bias and rule-of-thirds coverage than false positives. The lack of separation indicates that mid-level composition, when summarized globally, is not

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

a dominant factor in this task. One reason is ecological: persuasive imagery (ads, campaigns) commonly uses central placement and clear hierarchy, creating a population-wide center bias that collapses discriminability. A second is methodological: DeepGaze IIE-style saliency captures where attention concentrates, but it was trained on general visual attention, which may not be the same cognitive process as the attention used for judging persuasiveness. These findings indicate that global saliency measures alone offer limited explanatory power for persuasive reasoning, suggesting that composition effects become meaningful only when linked to semantically relevant regions of the image.

High-level (semantic) alignment dominates. Among all feature levels, high-level semantic alignment exerts the clearest and most interpretable influence on perceived persuasiveness. For human raters, the presence of the *message-relevant key object* in an image increases the odds of being judged persuasive by more than threefold (OR ≈ 3.28), reflecting strong semantic grounding between text and visuals. VLMs broadly follow this direction, achieving comparably strong though slightly weaker effects. In contrast, *human presence* has no significant influence for humans, yet models consistently overreact to it. Such over-sensitivity to other presence implies that models may conflate persuasive intent, revealing a systematic divergence in how humans and VLMs interpret semantic evidence.

Takeaway and implication. This discrepancy could contribute to the models' performance, highlighting the importance of aligning their reasoning processes with the cues and interpretations prioritized by humans. These observations suggest that improving VLM judgments requires tighter *grounding* of message nouns to detected objects and verifying their roles, coupled with a more robust *reasoning process* to evaluate *how* those objects contextually support the persuasive claim. We operationalize this in Section 5 and the knowledge-injection experiments by injecting a factor-aware context that emphasizes key-object grounding, aiming to examine how explicit grounding and diverse reasoning processes reveal that models often infer differently from humans despite observing the same visual information.

5 Cognitive Steering and VPF-aware Knowledge Injection

5.1 Task Description

To enhance the model's understanding of visual persuasion grounded in high-level factors, we apply human-identified cues—particularly key objects—to assess whether AI systems can reason about them as effectively as humans do. Based on this motivation, we propose four tasks designed to evaluate different levels of cognitive alignment. These tasks collectively test whether the models can (1) correctly recognize persuasive high-level factors (key objects) and (2) reason appropriately when such information is provided. All the details (e.g. prompt) are in Appendix C.

Cognitive Injection The system instruction is modified to make the model explicitly attend to *key object(s)* as a known persuasive cue. The instruction states the general principle that clear depiction of the core object tends to increase persuasiveness, without giving examples or labels. No other scaffolding or auxiliary context is provided; the task, inputs, and output space remain unchanged from baseline.

Knowledge-Conditioned Chain To guide the model's reasoning about key objects, a lightweight scaffold is introduced that frames them as *supporting—not determining*— cues. The model is asked to (1) extract candidate key objects, (2) assess visual–message alignment, and (3) give a final persuasiveness judgment. The chain encourages consideration of key objects while preventing over-reliance on them; the final decision is produced after the three steps.

Aligned Key-Object Context In this configuration, a compact structured context is injected, containing only high-level features from Section 4.1 for key-object candidates that semantically match the message. No information is added if no key-object match exists, in which case the input effectively reduces to baseline. The context provides object-level summaries (not labels or rationales).

Key-Object Rationale Building on the previous setting (Aligned Key-Object Context), we additionally supply a short, object-specific rationale that states *whether and why* the matched key object supports (or does not support) persuasiveness. This rationale is pre-generated by a generator (GPT-5)—which produces specific *per-object* rationales—before inference and then injected as context. It

Table 2: Results of key-object knowledge injection on VLMs performance.

Non-Aware Correct Key-Object												
Model	**Baseline**				**Cognitive Injection**				**Knowledge-Conditioned Chain**			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1 (Δ)	Acc.	Prec.	Rec.	F1 (Δ)
GPT-5	69.57	51.01	98.88	67.30	72.60	54.08	89.33	67.37 (↑0.07)	65.48	47.81	98.31	64.34 (↓2.96)
GPT-5-mini	60.14	44.13	97.19	60.70	60.68	44.56	98.88	61.43 (↑0.73)	63.35	46.30	98.31	62.95 (↑2.25)
Gemma3	63.88	46.33	88.76	60.89	61.03	44.38	91.01	59.67 (↓1.22)	39.15	33.98	97.75	50.43 (↓10.46)
Qwen2.5-VL	54.98	40.51	89.89	55.85	54.63	40.35	90.45	55.81 (↓0.04)	57.65	42.46	94.94	58.68 (↑2.83)
LLaVA 1.5	60.73	39.36	42.05	40.66	54.84	34.62	45.00	39.13 (↓1.53)	39.87	32.65	87.67	47.58 (↑6.92)
LLaVA-Next	59.17	31.65	25.00	27.94	62.16	38.13	29.94	33.54 (↑5.60)	42.11	33.89	87.65	48.88 (↑20.94)
Aware Correct Key-Object												
Model	**Aligned Key-Object Context**				**Key-Object Rationale (Agnostic)**				**Key-Object Rationale (Informed)**			
	Acc.	Prec.	Rec.	F1 (Δ)	Acc.	Prec.	Rec.	F1 (Δ)	Acc.	Prec.	Rec.	F1 (Δ)
GPT-5	69.93	51.44	90.45	65.58 (↓1.72)	69.40	50.89	96.07	66.54 (↓0.76)	85.05	68.08	99.44	80.82 (↑**13.52**)
GPT-5-mini	58.72	43.11	94.94	59.30 (↓1.40)	65.48	47.78	96.63	63.94 (↑3.24)	80.07	61.54	98.88	75.86 (↑**15.16**)
Gemma3	62.28	45.06	87.08	59.39 (↓1.50)	67.62	49.39	91.57	64.17 (↑3.28)	81.32	63.98	93.82	75.86 (↑**14.97**)
Qwen2.5-VL	53.02	39.56	91.57	55.25 (↓0.60)	61.74	44.79	89.33	59.66 (↑3.81)	77.40	59.14	92.70	72.21 (↑**16.36**)
LLaVA 1.5	53.28	38.46	59.52	46.73 (↑6.07)	66.67	50.00	60.87	54.90 (↑14.24)	69.47	57.14	59.57	58.33 (↑**17.67**)
LLaVA-Next	57.25	37.32	37.59	37.46 (↑9.52)	61.01	47.29	43.57	45.35 (↑17.41)	69.21	62.71	50.34	55.85 (↑**27.91**)

Notes. Results across configurations (%). Δ next to F1 shows absolute percentage-point change vs. the model's own baseline F1 scores (↑ / ↓ for direction).

also produces an *overall* rationale for the image's persuasiveness, used only for comparison rather than injection, as shown in Table 3 and details are in Appendix C.2. We evaluate two instantiations of the rationale generator: (i) *label-agnostic* (no access to the label) and (ii) *label-informed* (the generator knows the persuasiveness label when composing the rationale). Although the label-informed (oracle) generator unsurprisingly improves performance, we include it primarily to enable a controlled comparison of rationale similarity against the label-agnostic generator, and to serve as an upper-bound reference. Results reported in Table 2 use the label-informed variant or not.

5.2 Result

Non-Aware: instruction nudges are weak; scaffolds can over-amplify recall. *Cognitive Injection* changes F1 marginally on frontier models (GPT-5: ΔF1 $+0.07$; GPT-5-mini: $+0.73$), indicating that a generic "attend to key objects" cue is insufficient to counter the recall-oriented bias observed in Section 4.2. The lightweight knowledge-conditioned chain helps some non-frontier models by boosting recall without catastrophic precision loss (Qwen2.5-VL: ΔF1 $+2.83$), but it *hurts* stronger models in our setup (Gemma3: -10.46; GPT-5: -2.96), suggesting that unguided scaffolds can over-amplify the default tendency to predict *high* and miscalibrate precision.

Aware: features alone are not actionable; short object-grounded rationales work. Mirroring the human factor analysis in Table 1, we next asked whether making key-object information explicit helps. The *Aligned Key-Object Context* is neutral or slightly negative (GPT-5: ΔF1 -1.72; Gemma3: -1.50; Qwen2.5-VL: -0.60), implying that models often do not know *how* to use the features.

Table 3: Rationale similarity metrics

Set	BERTScore F1	ROUGE L	SBERT Cos.	BLEU
Overall	0.3550	0.7502	0.2715	0.0693
Per-Object	0.3984	0.7319	0.3128	0.0844
All (Both)	0.3813	0.7391	0.2966	0.0784

By contrast, *Key-Object Rationale (Agnostic)* improves F1 on models while tightening precision (GPT-5-mini: $+3.24$ with Prec. $+3.65$; Gemma3 & Qwen2.5-VL: $+3.28$ & $+3.81$ with both Prec. and Rec. up). Supplying an oracle-style *Key-Object Rationale (Informed)* yields the largest gains across all models, demonstrating that if a correct, concise, object-grounded rationale is available, models can align much more closely with human judgments.

The moderate similarity between *Agnostic* and *Informed* rationales in Table 3 underscores that even when models correctly identify key objects, they often fail to infer *how* those objects contribute to persuasiveness. Across all metrics, the overlap remains partial—ROUGE-L is high but mainly lexical, whereas semantic similarity (BERTScore-F1, SBERT-Cosine) stays modest and BLEU particularly low—indicating that the two rationale types diverge not in surface form but in underlying reasoning. This misalignment highlights that knowing *what* the key object is does not imply understanding *why* it matters, supporting our claim that explicit, object-grounded rationales are necessary for genuine persuasive reasoning.

6 Conclusion

In this work, we aimed to ask whether modern VLMs understand visual persuasion and evaluated them on a high-consensus dataset with a binary judgment task. Despite the task being binary, overall performance is modest and systematically skewed—frontier models show a recall-oriented tendency to over-predict *high*, while weaker open models lag across metrics. Grounded in cognitive psychology, our VPF taxonomy reveals that while high-level semantic alignment drives persuasiveness, state-of-the-art models respond to these cues differently from humans—over-attending to generic human presence while under-attending to message-relevant grounding. From the lens of cognitive steering and knowledge injection, both instruction-based and self-driven reasoning fail to resolve the gap: even when models are told what humans attend to, they lack the ability to reason why it persuades—highlighting a fundamental bottleneck in persuasive understanding.

7 Discussion

We note two main limitations in our study. First, our analysis focused on a high-agreement dataset, where persuasive cues are relatively clear to humans. While this allowed us to pinpoint fundamental failures, generalization to "low-agreement" data—where human interpretations widely vary—remains a necessary next step. Second, 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 works can not only extend the level-wise analysis of VPFs but also progress toward *merge-level* inference. At the low level, causal editing of color/contrast/texture can estimate sensitivity of persuasiveness under controlled perturbations. At the mid level, learning *persuasion-oriented* saliency—which encodes attention guidance, layout, and viewing paths—may strengthen the bridge from compositional cues to judgments. At the high level, the factor hierarchy should expand beyond objects/human to relations, scene context, affective cues, and text overlays. Collectively, these factors interact and merge, which also enables a new *merge-level* analysis. Practically, a dataset and metrics for generating *concise, object-grounded, label-agnostic* rationales are needed, alongside calibration protocols and finer-grained factor taxonomies. Finally, closing the loop by feeding these factors into text-to-image generation (factor-conditioned synthesis with automatic rationale generation) can enable a deployable pipeline that jointly improves understanding, decision, and generation.

Ethics Statement

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 VLM's 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.

Acknowledgment

The author would like to thank Prof. Seung-Hoon Na at UNIST NLP Lab for providing the computing resources necessary for this research.

References

[1] Junseo Kim, Jongwook Han, Dongmin Choi, Jongwook Yoon, Eun-Ju Lee, and Yohan Jo. PVP: An image dataset for personalized visual persuasion with persuasion strategies, viewer characteristics, and persuasiveness ratings. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19209–19237, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.942. URL https://aclanthology.org/2025.acl-long.942/.

[2] Richard E. Petty and John T. Cacioppo. *The Elaboration Likelihood Model of Persuasion*, pages 1–24. Springer New York, New York, NY, 1986. ISBN 978-1-4612-4964-1. doi: 10.1007/978-1-4612-4964-1_1. URL https://doi.org/10.1007/978-1-4612-4964-1_1.

[3] Seth M Noar, Marissa G Hall, Diane B Francis, Kurt M Ribisl, Jessica K Pepper, and Noel T Brewer. Pictorial cigarette pack warnings: a meta-analysis of experimental studies. *Tobacco Control*, 25(3):341–354, 2016. ISSN 0964-4563. doi: 10.1136/tobaccocontrol-2014-051978. URL https://tobaccocontrol.bmj.com/content/25/3/341.

[4] Eryn J. Newman and Norbert Schwarz. Misinformed by images: How images influence perceptions of truth and what can be done about it. *Current Opinion in Psychology*, 56:101778, 2024. ISSN 2352-250X. doi: https://doi.org/10.1016/j.copsyc.2023.101778. URL https://www.sciencedirect.com/science/article/pii/S2352250X23002233.

[5] Jiyoung Lee and Jiyoun Suk. Navigating the complexity of visual misinformation: Developing the visual misinformation processing model for visual-text misinformation dynamics. *Communication Theory*, 35 (4):238–249, 06 2025. ISSN 1468-2885. doi: 10.1093/ct/qtaf011. URL https://doi.org/10.1093/ct/qtaf011.

[6] Jungseock Joo, Weixin Li, Francis F. Steen, and Song-Chun Zhu. 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, June 2014. IEEE Computer Society. doi: 10.1109/CVPR.2014.35. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2014.35.

[7] Xuewei Wang, Yulong Zhang, Jemin Wang, Lili Li, Vaibhav Rajan, and Kevin Chang. Persuasion for good: Towards a personalized persuasive dialogue system. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5635–5649, 2019. doi: 10.18653/v1/P19-1566. URL https://aclanthology.org/P19-1566.

[8] Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. Empathetic persuasion: Reinforcing empathy and persuasiveness in dialogue systems. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 844–856, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.63. URL https://aclanthology.org/2022.findings-naacl.63/.

[9] S. C. Matz, J. D. Teeny, S. S. Vaid, et al. The potential of generative ai for personalized persuasion at scale. *Scientific Reports*, 14:4692, 2024. doi: 10.1038/s41598-024-53755-0. URL https://doi.org/10.1038/s41598-024-53755-0.

[10] H. Bai, J. G. Voelkel, S. Muldowney, et al. Llm-generated messages can persuade humans on policy issues. *Nature Communications*, 16:6037, 2025. doi: 10.1038/s41467-025-61345-5. URL https://doi.org/10.1038/s41467-025-61345-5.

[11] Daniel Chandler and Rod Munday. *A Dictionary of Media and Communication*. Oxford University Press, USA, 2011. ISBN 9780199568758. doi: 10.1093/acref/9780199568758.001.0001. eISBN: 9780191727979.

[12] Anthropic. Measuring model persuasiveness. https://www.anthropic.com/research/measuring-model-persuasiveness, 2024. Accessed 2025-10-17.

[13] Somesh Singh, Yaman K Singla, Harini SI, and Balaji Krishnamurthy. Measuring and improving persuasiveness of large language models, 2024. URL https://arxiv.org/abs/2410.02653.

[14] Lauren Labrecque and George Milne. Exciting red and competent blue: The importance of color in marketing. *Journal of The Academy of Marketing Science - J ACAD MARK SCI*, 40, 09 2011. doi: 10.1007/s11747-010-0245-y.

[15] Daniela Mapelli and Marlene Behrmann. The role of color in object recognition: Evidence from visual agnosia. *Neurocase*, 3(4):237–247, 1997. doi: 10.1080/13554799708405007. URL https://doi.org/10.1080/13554799708405007.

[16] Karl R Gegenfurtner and Jochem Rieger. Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology*, 10(13):805–808, 2000. ISSN 0960-9822. doi: https://doi.org/10.

1016/S0960-9822(00)00563-7. URL https://www.sciencedirect.com/science/article/pii/S0960982200005637.

[17] Francis M. Adams and Charles E. Osgood. A cross-cultural study of the affective meanings of color. *Journal of Cross-Cultural Psychology*, 4(2):135–156, 1973. doi: 10.1177/002202217300400201. URL https://doi.org/10.1177/002202217300400201.

[18] Lauren I. Labrecque, Vanessa M. Patrick, and George R. Milne. The marketers' prismatic palette: A review of color research and future directions. *Psychology & Marketing*, 30(2):187–202, 2013. doi: https://doi.org/10.1002/mar.20597. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mar.20597.

[19] Raymond P. Voss Jr., Ryan Corser, Michael McCormick, and John D. Jasper. Influencing health decision-making: A study of colour and message framing. *Psychology & Health*, 33(7):941–954, 2018. doi: 10.1080/08870446.2018.1453509. URL https://doi.org/10.1080/08870446.2018.1453509. PMID: 29667448.

[20] Katherine Armstrong, Adam S. Richards, and Kylie J. Boyd. Red-hot reactance: Color cues moderate the freedom threatening characteristics of health psas. *Health Communication*, 36(6):663–670, 2021. doi: 10.1080/10410236.2019.1700885. URL https://doi.org/10.1080/10410236.2019.1700885. PMID: 31818126.

[21] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2106–2113, 2009. doi: 10.1109/ICCV.2009.5459462.

[22] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. *CoRR*, abs/2105.12441, 2021. URL https://arxiv.org/abs/2105.12441.

[23] Alireza Hosseini, Amirhossein Kazerouni, Saeed Akhavan, Michael Brudno, and Babak Taati. Sum: Saliency unification through mamba for visual attention modeling. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1597–1607. IEEE, 2025.

[24] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32–32, 12 2008. ISSN 1534-7362. doi: 10.1167/8.7.32. URL https://doi.org/10.1167/8.7.32.

[25] Benjamin W. Tatler. 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, 11 2007. ISSN 1534-7362. doi: 10.1167/7.14.4. URL https://doi.org/10.1167/7.14.4.

[26] Jiwan Chung, Sungjae Lee, Minseo Kim, Seungju Han, Ashkan Yousefpour, Jack Hessel, and Youngjae Yu. Selective vision is the challenge for visual reasoning: A benchmark for visual argument understanding. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2423–2451, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.143. URL https://aclanthology.org/2024.emnlp-main.143/.

[27] Kiwon Seo. Meta-analysis on visual persuasion– does adding images to texts influence persuasion? *ATHENS JOURNAL OF MASS MEDIA AND COMMUNICATIONS*, 6:177–190, 06 2020. doi: 10.30958/ajmmc.6-3-3.

[28] Zesi Wang, Mohammad Rastegari, and Hannaneh Hajishirzi. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[29] Kanika Kalra, Bhargav Kurma, Silpa Vadakkeeveetil Sreelatha, Manasi Patwardhan, and Shirish Karande. Understanding advertisements with BERT. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7542–7547, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.674. URL https://aclanthology.org/2020.acl-main.674/.

[30] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[31] OpenAI. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf, August 2025.

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

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

[34] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STar: Bootstrapping reasoning with reasoning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_3ELRdg2sgI.

[35] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore, 2023. Association for Computational

Linguistics.

[36] Zhe Jiang, Jiaao Cui, Shizhe Diao, Xuhui Zhou, Tianyi Zhang, Ansong Ni, and Wayne Xin Zhao. Reft: Reasoning with reinforced fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–651, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.35.

[37] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=6LlR3A8fDxu.

[38] Michihiro Yasunaga, Xinyun Chen, Jure Leskovec, Percy Liang, et al. Large language models as analogical reasoners. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=AgDICX1h50.

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

[40] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[41] 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

[42] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

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

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

[45] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025.

[46] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025.

[47] David Hasler and Sabine E. Suesstrunk. Measuring colorfulness in natural images. In Bernice E. Rogowitz and Thrasyvoulos N. Pappas, editors, *Human Vision and Electronic Imaging VIII*, volume 5007, pages 87 – 95. International Society for Optics and Photonics, SPIE, 2003. doi: 10.1117/12.477378. URL https://doi.org/10.1117/12.477378.

[48] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2529310.

[49] 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. Simple open-vocabulary object detection with vision transformers. In *European Conference on Computer Vision (ECCV)*, 2022.

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

Table 4: Inter-annotator agreement statistics and the construction of the *robust* subset.

Metric	Original (0–10)	Banded (low/mid/high)
Overall Fleiss' κ	0.0233	0.0629
Mean per-item κ	0.023	0.063
Median per-item κ	0.003	0.011
Kappa level distribution [48] (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

Notes. 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' κ is reported overall as well as per-item distribution of agreement levels.

A Dataset Detail

Dataset and agreement filtering. We use the PVP dataset [1], where each message is paired with $\sim$50 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' κ score over the entire corpus is low for both schemes, but banding increases agreement and yields many "Almost perfect" items (Table 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.

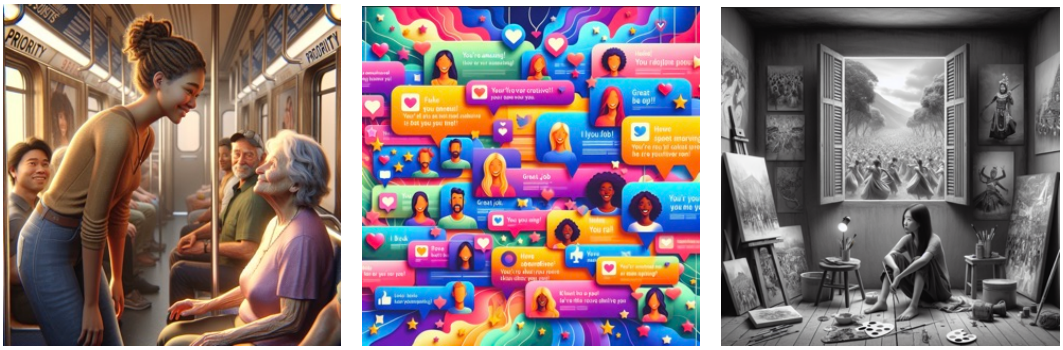
B Detailed VPF and Extracted Examples

This section provides the detailed definitions and equations for the VPF features described in Section 4.1.

B.1 Low-Level: Perceptual Features

We extracted two main features related to color to capture the immediate visual impression of an image.

- **Colorfulness:** To measure the intensity of colors, we used the metric $\hat{M}^{(3)}$ proposed by Hasler-Süsstrunk [47]. This metric is calculated by first transforming the sRGB color space into opponent color channels (red-green rg, yellow-blue yb). The colorfulness is then computed as a weighted sum of the mean (μ_{rgyb}) and standard deviation (σ_{rgyb}) of these



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

Figure 3: Illustrative low-level examples used in Appendix B.1

opponent color values.

$$rg = R - G \tag{1}$$

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

$$\sigma_{rgyb} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2}, \quad \mu_{rgyb} = \sqrt{\mu_{rg}^2 + \mu_{yb}^2} \tag{3}$$

$$\hat{M}^{(3)} = \sigma_{rgyb} + 0.3 \cdot \mu_{rgyb} \tag{4}$$

Interpretation: Typical reference values of $\hat{M}^{(3)}$ are: 0 (not colorful), 15 (slightly), 33 (moderately), 45 (averagely), 59 (quite), 82 (highly), 109 (extremely).

- **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_{color}, using the Shannon entropy formula, where a higher value indicates a more uniform distribution of diverse colors.

$$H_{\text{color}} = -\sum_i p_i \log_2 p_i \tag{5}$$

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

- **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^*:

$$\text{Brightness} = \frac{1}{HW} \sum_{x=1}^{H} \sum_{y=1}^{W} L^*(x, y). \tag{6}$$

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.

Example We provide illustrative examples of images with varying degrees of colorfulness in Figure 3, along with corresponding color metric values—specifically Hasler's colorfulness metric (`CF_Hasler`) and color entropy (`Entropy_bits`).

Color metrics for the examples in Figure 3.

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

B.2 Mid-Level: Compositional Features

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** [22].

- **Top-p Mass Area ($A@p$):** The fraction of image pixels required to accumulate top p mass of saliency. We use p=0.85; a smaller value indicates higher concentration.

$$A@p \;=\; \frac{k}{N}, \qquad k \;=\; \min\left\{ m : \sum_{\ell=1}^{m} S_{(\ell)} \geq p \right\}, \tag{7}$$

  where $S_{(\ell)}$ denotes saliency values sorted in descending order and N=HW the number of pixels.

- **Normalized Saliency Entropy (H_{sal}):** Shannon entropy of S normalized by its maximum possible value; lower is more concentrated.

$$H_{\text{sal}} \;=\; \frac{-\sum_{i,j} S_{i,j} \, \log S_{i,j}}{\log N}, \tag{8}$$

  where the logarithm base is arbitrary because it cancels in the ratio (we use natural log in implementation).

- **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 α=0.20.

$$\text{CBI} \;=\; \sum_{(i,j) \in D(c, r_c)} S_{i,j}, \qquad D(c, r) \;=\; \{(i,j) : \|(i,j) - c\|_2 \leq r\}. \tag{9}$$

- **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 β=0.10.

$$\text{T3} \;=\; \sum_{(i,j) \in \bigcup_{p \in \mathcal{P}} D(p, r_t)} S_{i,j}. \tag{10}$$

Example Figure 4 presents three representative mid-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.

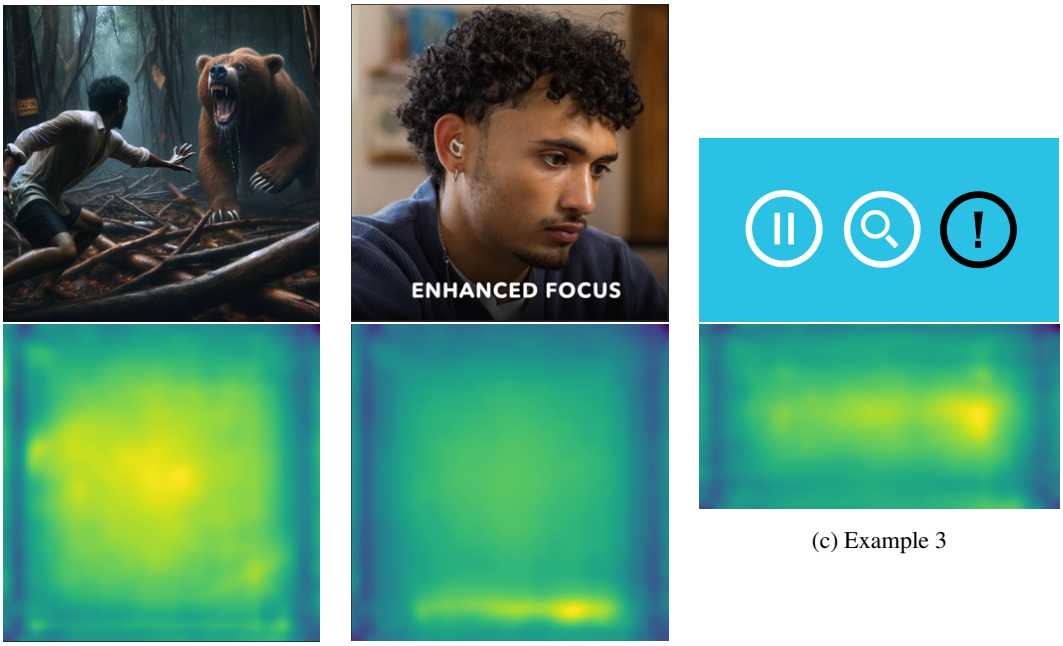
B.3 High-Level VPF Extraction Details

Models Used

- **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.
- **Open-Vocabulary Object Detection:** We employed the OWL-ViT v1 [49] 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.
- **Person Detection:** We used the pre-trained YOLOv8-nano [50] 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.

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

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



(a) Example 1 (b) Example 2 (c) Example 3

Figure 4: Illustrative examples used in Appendix B.2. Each example (top) is shown with its corresponding saliency map (bottom).

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

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

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

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).

We additionally computed average marginal effects (AME) on the probability scale:

$$\text{AME}_i = \frac{1}{n} \sum_{j=1}^{n} \frac{\partial \Pr(y_j = 1)}{\partial x_{ij}},$$

to interpret the expected change in predicted probability associated with each binary feature. All analyses were conducted using the `statsmodels` library in Python3.

This procedure allowed us to quantitatively assess how semantic alignment signals—specifically, the co-occurrence of mentioned and depicted entities—relate to perceptual judgments and model–human agreement.

Average Marginal Effects (AME) for High-Level Features As a complement to the odds–ratio analyses in this appendix, Table 5 reports average marginal effects (AME) on the probability scale for the two binary semantic features (Key Object Presence, Human Presence), computed from the logistic specifications in *Metrics and Analytical Procedure*. AMEs were obtained with `statsmodels`' `get_margeff` at the overall sample (HC3-robust inference; Benjamini–Hochberg FDR applied within each label/model block). Values are rounded to four significant figures and omit the leading zero for $x<1$. An AME can be read as the expected change in $\Pr(y{=}1)$ when the feature switches from 0 to 1.

Table 5: Average marginal effects (AME) on the probability scale for binary features (key-object presence, human presence).

Label / Model	AME(KeyObj)	AME(Human)
Human (Y)	.2393***	.04038
GPT-5	.2300***	.09738*
GPT-5-mini	.1954***	.09064*
Gemma3	.1607***	.1423***
Qwen2.5	.1892***	.1709***

Notes. AME computed at the overall sample with `statsmodels` `get_margeff`. *Reading guide.* Positive AMEs indicate that the corresponding semantic cue increases the probability of a "high" judgment; magnitudes are not directly comparable to ORs but provide probability-scale interpretability.

Clarification on outcome variables. Two outcome types are analyzed. (1) Human annotation (ground truth) — the binary persuasiveness label provided by human annotators; (2) Model prediction (label) — the corresponding binary decision produced by each VLM on the same image–message pairs. We previously denoted the latter as "GPT label," but to avoid confusion it is hereafter referred to as "model prediction (label)."

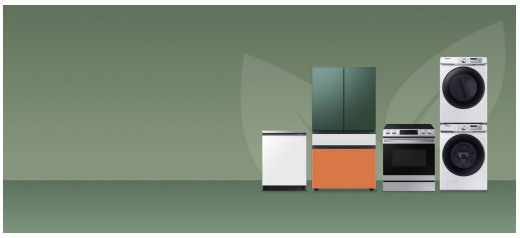
Table 6: Odds ratios (OR) with 95% CI from logistic regressions (robust SE). p_{FDR} = BH–FDR adjusted p-value.

Outcome	Model	Term	OR	95% CI (low, high)	p_{FDR}
Human annotation (ground truth)	–	KeyObj	3.2836	(2.2025, 4.8956)	$< .001$
	–	Human	1.2221	(0.7985, 1.8705)	.356
Model prediction (label)	GPT-5	KeyObj	2.868	(2.0076, 4.0971)	$< .001$
	GPT-5	Human	1.562	(1.0627, 2.2968)	.023
	GPT-5-mini	KeyObj	2.696	(1.8497, 3.9295)	$< .001$
	GPT-5-mini	Human	1.584	(1.0637, 2.3593)	.035
	Gemma3	KeyObj	2.041	(1.4362, 2.9015)	$< .001$
	Gemma3	Human	1.881	(1.2873, 2.7477)	.0016
	Qwen2.5	KeyObj	2.733	(1.8616, 4.0121)	$< .001$
	Qwen2.5	Human	2.479	(1.6646, 3.6926)	$< .001$

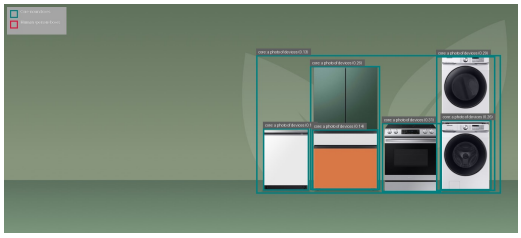
Extraction Example. To illustrate how high-level key objects are identified and mapped in our framework, Figure 5 presents two representative examples showing the original images and their corresponding object-detection overlays. The extracted original results for these samples are detailed in Json 1 and Json 2, demonstrating how each detected object is recorded in the JSON structure along with its bounding coordinates, confidence score, and source detector. These examples concretely show how persuasive key-object candidates (e.g., *devices*, *person*, *building*) are mapped from the visual input to structured indicators used in our analysis.

Json 1: Example A extracted result in Figure 5 (no human presence)


```
{
  "id": 413,
  "image_path": "/.../Remove_unnecessary_electronic_devices/27.jpg",
  "message": "Remove_unnecessary_electronic_devices",
  "indicators": {
    "core_noun_presence": {
      "binary": 1,
      "detail": {
        "nouns": ["devices", "electronic", "remove", ...],
        "matches": [{"noun": "devices", "count": 7}],
        "boxes": [
          {
```



(a) Example A — original



(b) Example A — detection

(c) Example B — original

(d) Example B — detection

Figure 5: High-level examples: each row shows a pair (original vs. object-detection overlay).


```
            "x_min": 1003.9299,
            "y_min": 315.0660,
            "x_max": 1161.0778,
            "y_max": 498.2654,
            "score": 0.314887,
            "label": "a photo of devices",
            "source": "owlvit"
          },
          // 6 more
        ]
      }
    },
    "human_presence": {
      "binary": 0,
      "detail": {"count": 0, "boxes": []}
    }
  }
}
```


Json 2: Example B extracted result in Figure 5 (human presence)


```
{
  "id": 153,
  "image_path": "/.../Use_recyclable_building_materials/1.jpg",
  "message": "Use_recyclable_building_materials",
  "indicators": {
    "core_noun_presence": {
      "binary": 1,
      "detail": {
        "nouns": ["building", "materials", "recyclable", "use"],
        "matches": [
          {"noun": "building", "count": 1},
          {"noun": "recyclable", "count": 1}
        ],
        "boxes": [
          {
            "x_min": -4.0028,
```

```
            "y_min": 175.7925,
            "x_max": 821.4266,
            "y_max": 657.3088,
            "score": 0.3029,
            "label": "a photo of building",
            "source": "owlvit"
          },
          // 1 more
        ]
      }
    },
    "human_presence": {
      "binary": 1,
      "detail": {
        "count": 14,
        "boxes": [
          {
            "x_min": 138.0458,
            "y_min": 825.5852,
            "x_max": 185.9777,
            "y_max": 982.7964,
            "score": 0.8731,
            "label": "person",
            "source": "yolo"
          },
          // 13 more
        ]
      }
    }
  }
}
```


C Prompts

All experiments were conducted on a single NVIDIA A6000 GPU. All models were loaded using 4-bit quantization, and most experimental runs completed within four hour. Also, model version details in Table 7.

C.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:
```

Table 7: Baseline models used for binary persuasiveness judgment. All models were evaluated with minimal reasoning capacity, except for the knowledge-conditioned chain (KCC) task, where reasoning capacity was set to medium.

Model	Version	Reasoning Effort
GPT-5 [31]	`gpt-5-2025-08-07`	minimal (medium for KCC)
GPT-5-mini	`gpt-5-mini-2025-08-07`	minimal (medium for KCC)
Gemma-3 [42]	`gemma-3-27b-it`	-
Qwen2.5-VL [41]	`Qwen2.5-VL-7B-Instruct`	-
LLaVA-1.5 [39]	`llava-1.5-13b-hf`	-
LLaVA-Next [40]	`llava-v1.6-mistral-7b-hf`	-

```
Message: {message}

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

### Response:
"""
```

C.2 Rationale Generator

We include two rationale-generation prompts that differ in whether the model is informed of the human persuasiveness label. The *agnostic* variant generates rationales without access to the label, while the *informed* variant conditions its reasoning on it. This comparison enables us to examine how awareness of persuasiveness influences the quality and explanatory alignment of the generated rationales.

Prompt 2: Key-object rationale generation (agnostic).

```
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 with a HUMAN persuasiveness label.

### 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>": "<...>"}},
  "rationale_overall": "<max 3 sentences>"
}}

### Inputs:
message: {message}
key_objects: {key_objects}
Return JSON only.

### Response:
"""
```

Prompt 3: Key-object rationale generation (informed).

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

User Prompt:
"""
Analyze the following image-message pair with a HUMAN persuasiveness label.

### Instructions:
1) Provide a concise overall rationale for why the pair is {label_upper}.
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>": "<...>"}},
  "rationale_overall": "<max 3 sentences>"
}}

### Inputs:
message: {message}
label: {label_lower}
key_objects: {key_objects}
Return JSON only.

### Response:
"""
```

Json 3: Agnostic rationale example for message "Use recyclable building materials" (without persuasiveness label).


```
{
  "analysis": {
    "persuasiveness_label": "high",
    "key_objects": ["building", "recyclable"],
    "rationale_overall": "The image is attractive and credible,
        showcasing an eco-friendly lifestyle that creates positive
        feelings toward sustainability. However, the message about
        using recyclable building materials is not directly evidenced
        -materials of the structure are not identified as recyclable,
         and attention shifts to general recycling. Overall
        persuasiveness is moderate: strong visuals, weak specificity.
        ",
    "per_object": {
      "building": "The lush, solar-topped house is central and
          evokes sustainable living, which supports the message. Yet
          there are no explicit cues that its materials are recycled
          or recyclable.",
      "recyclable": "Bright recycling bins with clear symbols are
          salient and signal 'recyclable,' reinforcing sustainability
          . However, they refer to household waste rather than
          construction materials, causing partial misalignment."
    },
    "skip_reason": null,
    "model": "gpt-5"
  }
}
```

Json 4: Informed rationale example for message "Use recyclable building materials" (with human persuasiveness label provided).

```
{
  "analysis": {
    "persuasiveness_label": "high",
    "key_objects": ["building", "recyclable"],
    "rationale_overall": "Strong message-image alignment: an
        attractive green building scene combined with visible
        recycling behaviors supports the idea of recyclable materials
        . High object clarity and emotional appeal from community
        participation add credibility. The visuals make the
        sustainability message concrete and actionable.",
    "per_object": {
      "building": "The eco-home with solar panels, green roof, and
          visible wood/stone textures highlights sustainable
          construction. Its appealing look links the message to
          desirable, credible outcomes implying recyclable materials.
          ",
      "recyclable": "Color-coded bins with recycling symbols and
          people sorting items make recyclability explicit. Their
          proximity to the house connects daily recycling to building
           choices, directly reinforcing the message."
    },
    "skip_reason": null,
    "model": "gpt-5"
  }
}
```

C.3 Cognitive Steering and VPF-aware Knowledge Injection Prompts

Prompt 4: Cognitive injection.

```
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. Particularly, pay strong attention to whether
    the message contains its key object(s), because clear depiction of the core
    object increases persuasiveness."

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:
"""
```

Prompt 5: 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 6: 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.
CRITICAL: Think critically about whether these objects actually support the
    message's persuasive intent. Do not upweight their presence by default.
    Consider relevance to the message, visual salience/clarity, emotional or
    safety impact, and consistency. If objects are irrelevant or contradict the
    message, this reduces persuasiveness.

### Input:
Message: {message}
```

```
Key Object: {key_object_json}

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

### Response:
"""
```

Prompt 7: Key-object rationale, only used when a key-object exists. The `{key_objects}` field contains the per-object rationales inserted verbatim from the `per_object` entries in the Json 3 or 4.

```
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:
"""
```