CAN MLLMS REASON ABOUT VISUAL PERSUASION? EVALUATING THE EFFICACY AND FAITHFULNESS OF REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Persuasive visuals play a central role in advertising, public communication, and online media, making it increasingly important to understand whether an image is persuasive and why. However, current Multimodal Large Language Models (MLLMs) have limited ability to reason about visual persuasion. Our analysis reveals that models often rely on a shortcut—selectively citing easily recognizable visual elements and treating their mere presence as evidence of persuasiveness—rather than reasoning over the diverse visual cues in an image to determine how and why they contribute to persuasiveness. To address this limitation, we propose (1) a fine-tuning approach that trains models on rationales reflecting diverse perspectives on visual persuasiveness, and (2) an evaluation framework that measures the faithfulness of models' rationales through three complementary metrics. Fine-tuning on multi-perspective rationales improves persuasiveness prediction and reasoning effectiveness. However, our evaluation framework reveals a discrepancy between prediction performance and rationale faithfulness, showing that higher prediction performance does not necessarily correspond to more faithful reasoning. These findings highlight the need to evaluate effectiveness and faithfulness separately and inform future directions for improving both the training and evaluation of visual persuasion reasoning.

1 INTRODUCTION

Visual persuasion, the use of images to influence cognition, beliefs, and behavioral intentions, plays a central role in modern communication (Chandler & Munday, 2011). Across public health campaigns, political messaging, and digital advertising, persuasive images serve as a medium for delivering messages that change the viewer's attitudes and behaviors. Advances in generative AI further increase the importance of this problem, as text-to-image models enable automated generation of tailored persuasive content at lower cost (Kim et al., 2025; Aghazadeh & Kovashka, 2025). As such content is generated and deployed at increasing scale, evaluating whether and why an image is persuasive becomes a problem of growing practical and societal significance.

Figure 1 (left) shows a target message, two images, and their persuasiveness labels. Despite the strong performance of MLLMs on many multimodal tasks, visual persuasiveness evaluation remains challenging for two main reasons. First, unlike factual image understanding, it requires interpretive reasoning over diverse factors, such as affective, symbolic, and compositional cues (Blair, 2011; Kjeldsen, 2015; McQuarrie & Mick, 1999; Phillips & McQuarrie, 2004), which can interact in complex ways to shape persuasiveness. Second, this reasoning is inherently non-unique: the persuasiveness of an image can vary with viewers' personality traits and values (Kim et al., 2025), allowing for multiple valid rationales for the same image and message and making it difficult to establish a consistent evaluation criterion.

Can current MLLMs reason about visual persuasion? Our analysis suggests they cannot (Sections 3 and 4), revealing two main problems. First, instructing models to generate rationales before predicting the persuasiveness of an image in relation to a target message does not reliably improve, and can even degrade prediction accuracy. Moreover, their reasoning often exhibits a message-relevance shortcut: identifying visual elements related to the target message and treating the mere presence of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

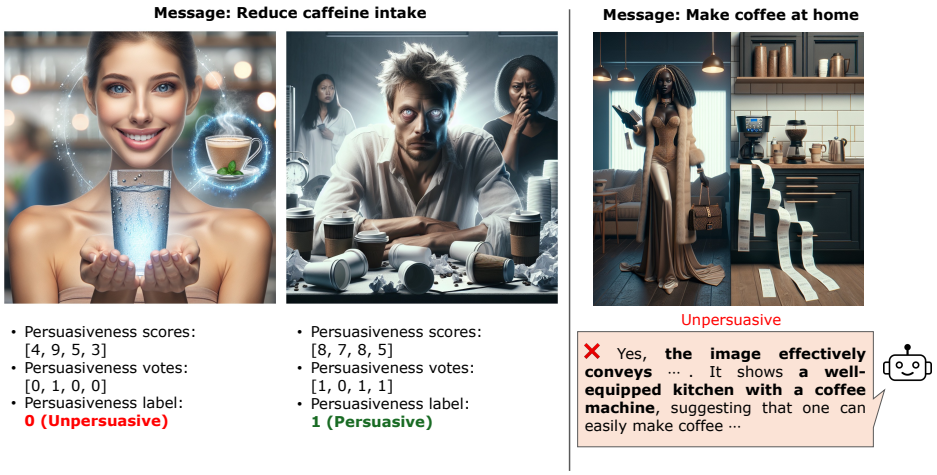


Figure 1: **The PVP dataset and a failure mode of MLLMs. (Left)** Example image–message pairs from the PVP dataset. *Persuasiveness scores* are the 0–10 ratings assigned by individual annotators, while *persuasiveness votes* are binary votes converted based on each annotator's score distribution. The *persuasiveness label* is determined by aggregating these binary votes. **(Right)** A *message-relevance shortcut*, where the model treats the presence of message-related visual elements as evidence of persuasiveness.

such elements as sufficient evidence of persuasiveness. For example, as shown in Figure 1 (right), a model treats the presence of a coffee machine as evidence that an image is persuasive for the message "Make coffee at home", despite the image being labeled unpersuasive by humans. These findings highlight two challenges: (1) improving the reasoning capabilities of MLLMs for visual persuasion, and (2) evaluating the faithfulness of generated rationales to both the input images and the model's final decisions.

To address the first gap, we construct a rationale dataset from large vision-language teacher models. We improve rationale diversity by introducing dual-axis prompts that vary along *evidence polarity* (*including* vs. *excluding* opposing evidence for persuasiveness) and *visual granularity* (*global* vs. *local*). This approach also mitigates the teacher-model's biases in reasoning patterns. Fine-tuning on the resulting multi-perspective rationales consistently improves persuasiveness prediction over training-free and label-only objective baselines.

For the second gap, we introduce a novel evaluation framework for reasoning faithfulness in visual persuasion, assessing rationales along three dimensions: *rationale-to-image groundedness*, whether the claims in the rationale are supported by visual evidence in the image; *rationale-to-decision consistency*, whether the model's final decision is consistent with the rationale's overall assessment; and *rationale-to-decision sensitivity*, whether the model's decision is sensitive to changes in the visual evidence identified by the rationale. This framework reveals that fine-tuning on the multi-perspective rationales improves rationale-to-image groundedness and rationale-to-decision consistency as the training rationales specify diverse visual components that support or counter image persuasiveness, while showing a trade-off in rationale-to-decision sensitivity. Overall, our work lays a foundation for training approaches that jointly improve prediction performance and reasoning faithfulness in visual persuasiveness prediction.

2 Related Work

2.1 Visual Persuasion and Visual Rhetoric

Computational work on visual persuasion has progressed from inferring communicative intent from images of political figures (Joo et al., 2014), to mining image persuasiveness in multimodal social

media posts (Liu et al., 2022), and collecting large-scale persuasiveness ratings (Kim et al., 2025). These efforts treat persuasiveness as a prediction target without examining reasoning behind those judgments. Theoretical work in visual rhetoric establishes, however, that persuasive images operate through affect, symbolism, and argument structure (Blair, 2011; Kjeldsen, 2015; McQuarrie & Mick, 1999; Phillips & McQuarrie, 2004), with meaning constructed across multiple levels of visual perception (Navon, 1977; Oliva & Torralba, 2006). Correct prediction alone cannot reveal whether a model's judgment is grounded in a coherent visual argument, which our work aims to measure.

2.2 Rationale Supervision

A growing body of work has shown that fine-tuning language models on chain-of-thought rationales generated by larger teacher models, rather than output labels alone, improves task performance and generalization (Magister et al., 2023; Ho et al., 2023; Shridhar et al., 2023). This paradigm has been extended to MLLMs, where supervising intermediate reasoning over visual inputs further improves performance on complex multimodal tasks (Zhang et al., 2023; Shao et al., 2024a). These methods, however, largely treat label correctness as sufficient supervision, leaving the quality of the reasoning process itself unmeasured. In visual persuasion, where persuasive judgments admit multiple valid interpretations, correct label prediction alone does not guarantee that the model's rationale faithfully reflects the basis of its judgment.

2.3 Rationale Faithfulness Evaluation

In NLP systems, faithfulness refers to how accurately an explanation reflects a model's reasoning process (Jacovi & Goldberg, 2020). Prior work has studied faithfulness along dimensions, including the sufficiency and comprehensiveness of extractive rationales (DeYoung et al., 2020), prediction sensitivity to reasoning-step perturbations (Lanham et al., 2023), and the causal influence of intermediate reasoning steps on final outcomes (Paul et al., 2024). These approaches, however, assume settings where explicit evidence, structured reasoning steps, or well-defined decision spaces provide anchors for evaluating reasoning faithfulness. Extending these criteria to visual persuasion is nontrivial, as rationales may rely on subjective, symbolic, and affective interpretations of visual cues, need not follow a fixed sequence of reasoning steps, and admit multiple plausible reasoning paths for the same image and message. To the best of our knowledge, we introduce the first rationale evaluation protocol for this setting, assessing reasoning quality along rationale-to-decision consistency, rationale-to-image groundedness, and rationale-to-decision sensitivity.

3 Dual-Axis Rationale Training for Visual Persuasiveness Prediction

3.1 Motivation

As a preliminary analysis, we examined whether MLLMs' reasoning improves their persuasiveness prediction. Table 1 compares instructing MLLMs to predict directly (Base) with instructing them to generate a rationale before predicting (Base-Reasoning). Base MLLMs achieve only modest performance under direct prediction, and rationale-first prediction does not consistently improve over this baseline. To understand this limitation, we analyzed rationales generated by the base models and found a message-relevance shortcut: MLLMs often treat message-related visual elements as sufficient evidence of persuasiveness. Figure 1 (right) illustrates this: for the message "Make coffee at home", the model cites the presence of a coffee machine to justify its prediction of 'persuasive', even though the ground-truth label is 'unpersuasive'.

This limitation is thus not merely a visual recognition failure, but an *evidence-use failure*: MLLMs rely on message-relevant cues while underweighting evidence against persuasiveness, such as distracting objects that draw attention from the intended message, or cues too weak to make the target action compelling. Their rationales therefore justify a positive prediction with selectively supportive cues, rather than weighing supporting and opposing evidence for an image-level judgment.

This also motivates rationale supervision beyond the model's own trajectories. Since online trajectories are sampled from the model's current reasoning policy, they expose the model to a limited set of rationale patterns. If the base model already equates message relevance with persuasiveness,

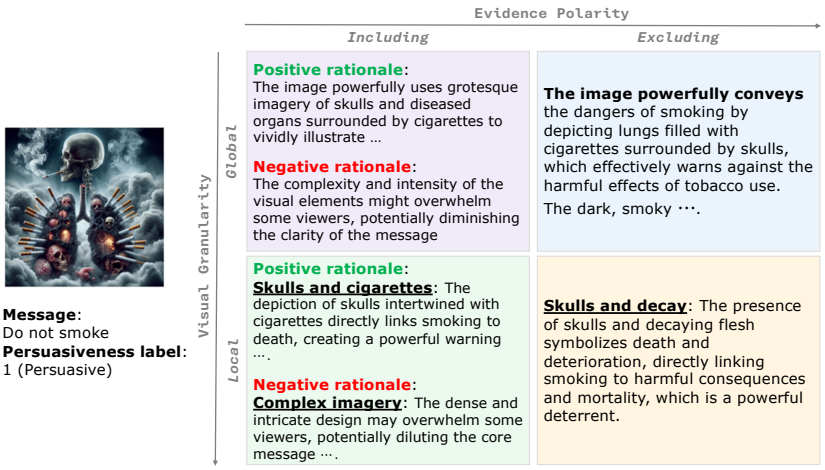


Figure 2: **Dual-axis rationale design.** Four rationale types defined by two axes: evidence polarity (including vs. excluding opposing evidence) and visual granularity (global vs. local). Given an image, its message ("Do not smoke"), and its persuasiveness label, each prompt type elicits a rationale with different characteristics.

optimizing over such trajectories may reinforce this shortcut rather than teach reasoning patterns outside its current capability.

3.2 Dual-Axis Rationale Design: Evidence Polarity and Visual Granularity

To diversify reasoning perspectives used for supervision, we design dual-axis prompts that vary the type and level of visual evidence considered in the rationale. The first axis, *evidence polarity*, controls whether the rationale considers only supporting evidence for persuasiveness or both supporting and opposing evidence. The second axis, *visual granularity*, controls whether the rationale focuses on the overall scene or on specific visual elements. Figure 2 summarizes the resulting prompt design.

Evidence polarity specifies whether a rationale uses only evidence supporting the persuasiveness decision or also considers evidence that weakens it. This axis is motivated by cognitive psychology work showing that robust judgment requires considering not only belief-consistent evidence, but also conflicting cues and alternative interpretations (Baron, 2005; Stanovich & West, 1997; Nickerson, 1998; Stanovich et al., 2013). Hence, we define two prompt types for guiding the model to *include* or *exclude* opposing evidence for persuasiveness.

Visual granularity specifies whether a rationale evaluates the image as a whole or focuses on specific visual evidence. This axis is motivated by work in visual cognition showing that perception operates across global and local levels, from scene structure to individual elements (Navon, 1977; Oliva & Torralba, 2006). We define two prompt types: *global* and *local*. Global prompts guide the model to evaluate the scene, composition, and communicative context, whereas local prompts guide it to examine specific objects, actions, text, and visual details relevant to the persuasiveness decision.

3.3 Rationale Dataset Construction

To construct rationales for training MLLMs, we use the PVP dataset (Kim et al., 2025), which consists of both AI-generated and web-sourced images. We focus on the AI-generated subset for licensing and reproducibility reasons, and many web-sourced images are misaligned with the target messages. The AI-generated subset of PVP contains 14,161 images paired with 596 messages across 20 topics, with each image–message pair rated for persuasiveness on a 0–10 scale by four annotators.

We formulate persuasiveness evaluation as a binary classification task of persuasive vs. unpersuasive (Tan et al., 2016; Liu et al., 2022; 2023). Although persuasiveness is viewer-dependent, as an initial study of visual persuasiveness reasoning, we focus on images where annotators exhibit sufficient agreement. This provides more stable labels that are less sensitive to idiosyncratic viewer pref-

erences and allows us to investigate models' reasoning about visual components without introducing more complex viewer factors. To construct the binary labels, we first account for annotator-specific rating scales by converting each annotator's raw score into a binary vote using the first and third quartiles of their score distribution (Figure 1 left). Based on these binary votes, we retain image–message pairs as persuasive or unpersuasive only when at least 75% of votes agree, discarding the remaining pairs. The resulting dataset has a mean vote agreement of 0.875. Finally, for each of the five random seeds, we split the retained pairs by message while maintaining label balance. This ensures that no message overlaps between the training and test sets, so that evaluation measures generalization to unseen messages. Averaged over the 5 seeds, the training split contains 823 examples and the held-out test split contains 206 examples. Further details on dataset reconstruction are provided in Appendix A.

For each image–message–label triple in the training split, we apply all seven prompt templates from Appendix B, covering different prompting configurations along the two axes of evidence polarity and visual granularity. These templates are designed to elicit rationales from different perspectives by varying whether opposing evidence is considered and whether the rationale focuses on global or local visual evidence. We construct four teacher-generated rationale sets: *Qwen*, generated by Qwen2.5-VL-72B-Instruct (Team, 2025); *Phi*, generated by Phi-4-reasoning-vision-15B (Aneja et al., 2026); *Gemma*, generated by Gemma-3-27B-IT (Team et al., 2025); and *LLaVA*, generated by LLaVA-1.6-34B (Liu et al., 2024). After filtering out rationales that do not follow the required output format, each rationale set contains roughly 5,700 instances.

3.4 Experiments on Dual-Axis Rationale Supervision

3.4.1 Experimental Setup

We fine-tune student models—Qwen2.5-VL-7B-Instruct, Phi-3.5-Vision-Instruct, Gemma-3-4B-IT, and LLaVA-1.6-7B—on the *Qwen*, *Phi*, *Gemma*, and *LLaVA* rationale sets, respectively. Each student is paired with a teacher from the same model family to mitigate feature-space misalignment (Hao et al., 2023; Boizard et al., 2024). We compare our *Reasoning-SFT* against five baselines grouped by training objective. *Training-Free* baselines include Base, which directly predicts the label, and Base-Reasoning, which generates a rationale before prediction. *Answer-Only Objective* methods include SFT, trained with supervised next-token prediction on labels, and GRPO-Ans (Shao et al., 2024b), which uses a final-answer reward while applying the loss only to answer tokens. *Answer + Rationale Objective* methods include GRPO, which uses the same final-answer reward but applies the loss to the full rationale–answer sequence, and our Reasoning-SFT, which uses supervised next-token prediction on teacher-generated rationales and labels. The prompt templates used for each setup and implementation details are provided in Appendix C and Appendix D, respectively.

3.4.2 Results

Table 1 reports persuasiveness classification performance on the PVP test split over five random seeds. Reasoning-SFT achieves the highest balanced accuracy on three of the four student models, improving over the strongest baseline by 0.5–3.6 percentage points (pp); on LLaVA-1.6-7B, it remains competitive with answer-only SFT. As an additional analysis, we test whether the benefit of Reasoning-SFT can be reproduced by simply prompting an answer-only SFT model to generate a rationale before its prediction. Across five seeds on Qwen2.5-VL-7B-Instruct, the strongest answer-only SFT model, the generated rationales tend to describe the image favorably regardless of the gold label, and the model predicts the 'persuasive' class for 68.1% of test examples, even though only about half are persuasive. As a result, balanced accuracy drops to 69.1%, compared to 74.6% when the same model directly predicts the label without a rationale. This degradation indicates that answer-only supervision does not readily transfer to effective rationale generation. In contrast, Reasoning-SFT is explicitly trained to reason before prediction and can generate rationales while maintaining classification performance comparable to or better than answer-only training.

Generating rationales without supervising them, by contrast, provides no reliable benefit. Prompting the base model to reason before answering (Base-Reasoning) even degrades balanced accuracy on three of the four student models, by up to 8.3 pp, suggesting that base MLLMs cannot effectively reason about persuasiveness. Both GRPO variants also fall below label-only SFT on all four models. GRPO—which applies reward over the full rationale-and-answer sequence—underperforms

Table 1: Persuasiveness classification performance (%) of student models on the PVP test split. All numbers are averaged over five random seeds. *Our method* (Reasoning-SFT) fine-tunes the student using supervision on teacher-generated rationales from diverse reasoning perspectives. BA: balanced accuracy; P: precision; R: recall. High recall with low precision indicates over-prediction of the 'persuasive' class. **Bold** indicates the best BA and F1 within each column. Detailed results with standard deviations across the five random seeds are provided in Appendix E.1.

	Qwen2.5-VL-7B				Phi-3.5-Vision				Gemma-3-4B				LLaVA-1.6-7B			
Setup	**BA**	**F1**	**P**	**R**	**BA**	**F1**	**P**	**R**	**BA**	**F1**	**P**	**R**	**BA**	**F1**	**P**	**R**
Training-Free:																
Base	63.2	68.7	59.6	81.3	65.2	69.9	61.4	81.1	61.7	69.9	57.4	89.6	50.6	66.7	50.1	99.8
Base-Reasoning	61.9	68.9	57.9	85.0	56.9	64.6	54.7	77.1	55.2	63.9	53.4	79.6	52.0	66.1	50.9	94.3
Answer-Only Objective:																
SFT	74.6	75.0	75.0	75.3	73.7	**75.6**	70.5	81.7	72.1	71.8	72.2	71.7	**72.9**	**74.6**	70.3	79.9
GRPO-Ans	70.8	71.3	70.0	72.8	67.5	66.6	74.3	60.5	71.5	73.4	68.5	79.4	53.2	67.2	51.7	96.5
Answer + Rationale Objective:																
GRPO	69.2	73.2	64.4	85.0	55.3	67.6	52.7	94.2	66.7	72.7	61.5	89.0	56.1	68.5	53.2	96.3
Reasoning-SFT	**75.1**	**75.1**	75.7	74.7	**74.8**	74.2	76.3	72.2	**75.7**	**75.3**	76.7	74.3	72.5	72.5	72.7	72.6

GRPO-Ans on three of the four models and shows signs of prediction collapse on most of them. We hypothesize that this collapse stems from the base models' prior bias toward the 'persuasive' class combined with outcome-based rewards: they achieve persuasive-class recall of 81.1–99.8% but precision of only 50.1–61.4%, indicating that they overpredict the 'persuasive' class. Because the reward depends only on final-answer correctness, it may reinforce rationale–answer trajectories that follow this bias whenever they yield a correct prediction, regardless of rationale quality. This can further encourage overprediction of 'persuasive', resulting in lower balanced accuracy. In Section 4, we further investigate this behavior by evaluating the faithfulness of the generated rationales.

4 Faithfulness Evaluation for Persuasiveness Rationales

Reasoning-SFT with dual-axis rationales improves visual persuasiveness prediction, and its improvement over label-only SFT suggests that rationale supervision provides benefits beyond label supervision. However, the generated rationales may cite relevant visual evidence without influencing the model's final decision. This concern is particularly relevant to visual persuasion, where models must not only identify message-relevant visual elements but also explain how they contribute to whether the image is persuasive.

Assessing whether rationales faithfully provide such explanations is challenging because there is no single faithful rationale that can serve as the ground truth—the same image, target message, and label admit multiple valid rationales. Existing faithfulness evaluation methods are also not well suited to our setting, since visual persuasion often requires interpreting visual evidence rather than simply identifying it. We therefore introduce a new framework for evaluating the faithfulness of generated rationales in visual persuasiveness along three complementary dimensions: whether they are grounded in the image, whether they support the predicted label, and whether changing the cited visual evidence affects the model's prediction.

4.1 Evaluation Metrics

Figure 3 summarizes the evaluation pipelines for the three faithfulness metrics. All metrics are computed as $\{0, 1\}$ scores, where 1 indicates a faithful rationale along the corresponding dimension.

4.1.1 Rationale-to-Image Groundedness

Rationale-to-image groundedness measures whether the rationale is grounded in visual evidence from the image. This metric identifies rationales that describe or infer visual evidence that is not supported by the image. Evaluating groundedness is challenging in visual persuasion because the relevant evidence is not limited to objects or regions. Rationales often refer to emotional tone,

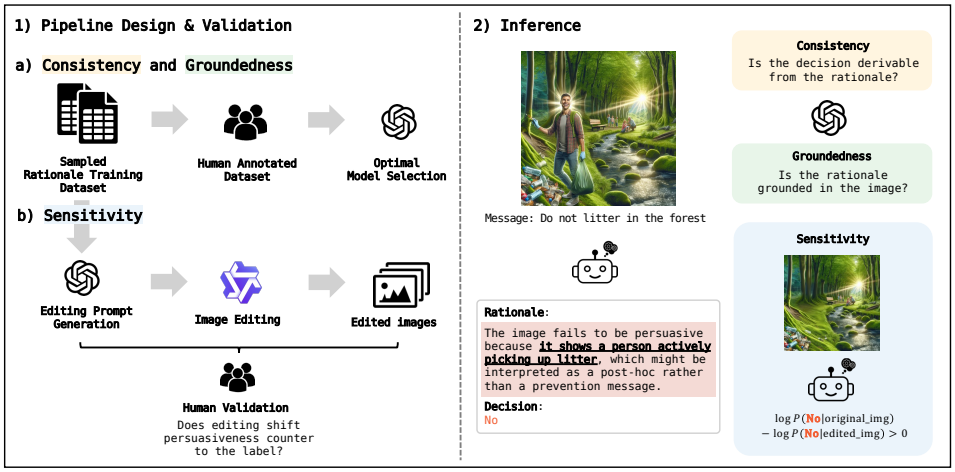


Figure 3: **Overview of faithfulness evaluation pipelines.** (Left) We validate judge-based pipelines for *rationale-to-decision consistency*, and *rationale-to-image groundedness* against human annotations, and the editing pipeline for *rationale-to-decision sensitivity*. (Right) At inference, GPT-5 judges consistency and groundedness, while sensitivity is measured by the log-probability gap between original and edited images for the predicted label.

symbolism, composition, and image–message alignment, which are difficult to capture with existing object-centric grounding metrics (Park et al., 2018; Selvaraju et al., 2019; Salewski et al., 2020).

Accordingly, we considered five methods: CLIP-based similarity with text or image inputs, GPT-5-based atomic-fact verification with and without calibration, and direct GPT-5 prompting (Appendix G.2). To assess how well their judgments align with human judgments, we sampled 10% of the training data, asked three annotators to assess whether each rationale is grounded in the image, and took the majority votes. The annotations show substantial inter-annotator agreement (Fleiss' $\kappa = 0.680$). Based on this evaluation, we select direct GPT-5 prompting as the final judge.

Although direct GPT-5 prompting performs best, qualitative analysis shows that it judges fine-grained visual claims conservatively: e.g., "There is a phone" is judged grounded but "There is an old phone" is not. Nevertheless, 93.6% of GPT-5's positive judgments match the human labels, and it identifies 88.0% of truly ungrounded rationales. After manually excluding these conservative cases, GPT-5 achieves a balanced accuracy from 57.1% to 63.7% and macro-F1 from 54.9% to 56.9%. We therefore attribute most of GPT-5's errors to overly conservative judgments on grounded rationales rather than failures to detect grounding errors. We assign a groundedness score of 1 when GPT-5 judges the rationale as grounded in the corresponding image, and 0 otherwise.

4.1.2 Rationale-to-Decision Consistency

Rationale-to-decision consistency measures whether the predicted persuasiveness decision is derivable from the generated rationale. This metric tests whether the rationale provides sufficient evidence for the model's predicted label. To assess how well consistency judgments align with human judgments, we sampled 10% of the training data and asked three annotators to judge whether each rationale is sufficient to derive the decision. We obtained the human labels by majority vote, with near-perfect inter-annotator agreement (Fleiss' $\kappa = 0.996$). We then evaluated GPT-5, Qwen3-32B, Qwen3-8B, and an NLI-based DeBERTa model as candidate judges against these human labels. GPT-5 achieved the highest performance against the human labels, attaining 99.6% for both balanced accuracy and F1 (Appendix G.3). We therefore adopt GPT-5 as the final judge and assign a score of 1 when it determines that the decision is derivable from the rationale, and 0 otherwise.

4.1.3 Rationale-to-Decision Sensitivity

Rationale-to-decision sensitivity measures whether the model's decision is sensitive to visual evidence cited in the rationale. Whereas consistency and groundedness assess properties of the rationale

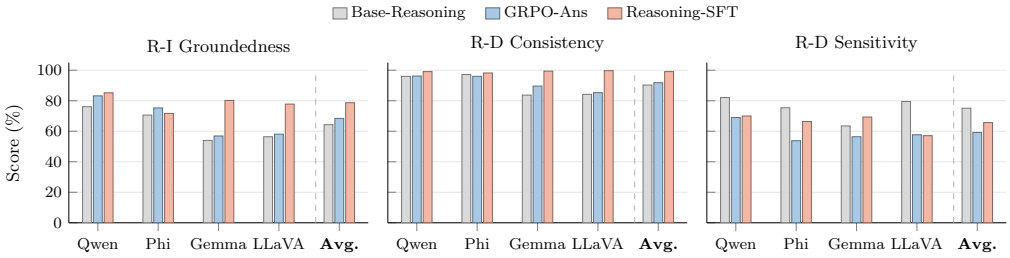


Figure 4: Rationale faithfulness (%) of each training setup across four student models and their average, over five random seeds. Reasoning-SFT yields the highest R-I groundedness and R-D consistency, while Base-Reasoning retains the highest R-D sensitivity. Exact mean ± standard deviation values are reported in Table 15 (Appendix I.1).

itself, sensitivity tests whether intervening on rationale-cited evidence changes the model's decision behavior. This metric follows perturbation-based faithfulness evaluation, which tests whether removing or altering the evidence identified in an explanation changes the model's prediction (DeYoung et al., 2020; Hooker et al., 2019; Atanasova et al., 2023). Because visual persuasiveness rationales describe semantic evidence rather than pixel-level masks, we implement this idea through rationale-conditioned counterfactual image editing.

Given an original image I, target message m, rationale r, and predicted label $\hat{y}$, we proceed in three steps. First, GPT-5 converts r into an editing instruction that weakens the rationale-cited evidence supporting $\hat{y}$, and Qwen-Image-Edit (Wu et al., 2025) generates the edited image I'. Second, we feed I' and m back into the model using the same prompt. Third, we compute $\Delta_{\text{sens}} = \log P(\hat{y} \mid I, m) - \log P(\hat{y} \mid I', m)$ and assign a binary score of 1 if $\Delta_{\text{sens}} > 0$ and 0 otherwise, so that a score of 1 indicates that confidence in the original prediction drops after the edit.

Since sensitivity depends on the quality of the image editing, three annotators validated the editing pipeline, showing substantial agreement (Fleiss' $\kappa = 0.693$) and judging 74.9% of edited images as successfully shifting persuasiveness against the predicted label (Appendix G.7). Appendix H reports control experiments with random and persuasiveness-preserving edits, verifying that sensitivity reflects the influence of rationale-cited evidence rather than a general response to editing.

4.2 Rationale Faithfulness Results

We evaluate rationale faithfulness for three setups that generate rationales: Base-Reasoning (no training), GRPO-Ans (answer-only reward), and Reasoning-SFT (rationale supervision). We use GRPO-Ans as the RL baseline because it achieves higher balanced accuracy than GRPO on average (65.8% vs. 61.8%; Table 1). Figure 4 presents the quantitative faithfulness results across the three metrics. Qualitative analyses are also provided in Appendix I.2.

4.2.1 Rationale-to-Image Groundedness

For rationale-to-image groundedness, Base-Reasoning achieves the lowest average score, while GRPO-Ans improves by only about 4 pp. Because GRPO-Ans receives reward only for the final answer, rationales that lead to correct predictions can be reinforced regardless of whether they are grounded in the image. Reasoning-SFT improves groundedness by approximately 15 pp over Base-Reasoning and 10 pp over GRPO-Ans. Since teacher rationales specify various visual components from diverse reasoning perspectives, Reasoning-SFT may encourage the student to attend more closely to image-specific information. Such supervision may help the student incorporate these task-relevant visual details into its own rationales, resulting in stronger grounding in the image.

To investigate the suboptimal groundedness scores across all three setups (60–80%), we randomly sampled from each of the four Reasoning-SFT training splits the same number of examples as in the test set and evaluated their consistency and groundedness. The sampled rationales show near-perfect consistency (98.1–100), whereas groundedness is substantially lower (79.9–85.6). This gap suggests that image-grounded rationale generation is inherently more challenging than achieving ra-

tionale–decision consistency and that using more grounded teachers may be able to improve student models' groundedness.

4.2.2 RATIONALE-TO-DECISION CONSISTENCY

For rationale-to-decision consistency, Reasoning-SFT achieves the highest average score, outperforming Base-Reasoning by approximately 9 pp and GRPO-Ans by 7 pp. In the Reasoning-SFT training data, each teacher rationale is generated based on the ground-truth label, resulting in high consistency between the rationale and the label. Therefore, training the student model to generate such a rationale and then the final answer provides direct supervision for producing answers that are consistent with the model's own reasoning. In contrast, Base-Reasoning receives no training signal that explicitly enforces consistency between the rationale and the final answer. GRPO-Ans likewise provides reward only for the final answer, so a sample can receive the same reward even when the rationale is inconsistent with the answer, as long as the answer itself is correct. As a result, GRPO-Ans lacks an explicit signal for penalizing rationale–answer inconsistency.

4.2.3 RATIONALE-TO-DECISION SENSITIVITY

For rationale-to-decision sensitivity, Reasoning-SFT generally retains sensitivity to the visual evidence cited in the rationale, although LLaVA shows a notable decrease. One possible explanation is that rationale supervision has a dual effect. When teachers are instructed to specify diverse visual evidence in their rationales, this naturally introduces less critical visual components along with more significant ones. As a result, training the student on these rationales may encourage the model to cite not only decision-sensitive evidence but also less decision-sensitive evidence. Consequently, such teacher rationales may introduce a trade-off between image groundedness and decision sensitivity, and improving both properties in rationales would be a fruitful direction.

5 FUTURE DIRECTIONS

Our findings suggest two directions for future work. First, groundedness and sensitivity could be further improved through higher-quality rationale data and targeted training objectives. As discussed above, the training rationales themselves exhibit relatively low groundedness, and the groundedness achieved by Reasoning-SFT appears to be partly constrained by that of its training data. Constructing more strongly grounded rationales may therefore provide more effective supervision for improving rationale groundedness. Sensitivity is not directly supervised during training, and both GRPO-Ans and Reasoning-SFT show lower sensitivity than Base-Reasoning. This motivates intervention-based objectives that explicitly encourage sensitivity to rationale-cited visual evidence.

Second, persuasiveness classification performance and rationale faithfulness should be evaluated separately. The discrepancy between the classification results in Table 1 and the faithfulness results in Figure 4 indicates that higher classification performance does not necessarily correspond to greater rationale faithfulness.

6 CONCLUSION

MLLMs often cite easily recognizable visual elements and treat their mere presence as evidence of persuasiveness, which can lead to incorrect predictions and raises questions about whether their rationales faithfully reflect the visual evidence. In this paper, we investigated both the effectiveness and faithfulness of MLLM reasoning for visual persuasion. For effectiveness, we proposed dual-axis rationale supervision, varying prompts along evidence polarity and visual granularity to provide diverse reasoning signals for fine-tuning. This approach improves the reasoning effectiveness, with higher persuasiveness classification performance than the baselines. For faithfulness, we further introduced, to our knowledge, the first rationale faithfulness evaluation framework for visual persuasion. Applying this framework reveals a discrepancy between persuasiveness classification performance and rationale faithfulness, highlighting the need to evaluate them separately. Based on these findings, we outline future directions for improving both the training and evaluation of effective and faithful visual persuasion reasoning.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic rationale data and to assist with translation. We have not used generative AI tools to develop theoretical models or conceptual frameworks, propose or refine hypotheses, design research methodology or experiments, implement methods, clean or reformat datasets, or interpret results. Formulating mathematical claims, assisting with proofs, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to edit the paper to improve readability. We have reviewed all AI-assisted work. The synthetic rationales were manually examined by the authors for quality before being used in our experiments, and all translated and edited text was reviewed by the authors to ensure that its meaning and technical content were preserved. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves three human evaluation studies for validating the faithfulness metrics. The groundedness and consistency evaluations were conducted by three of the authors, while the sensitivity evaluation involved volunteers who assessed whether the counterfactual image edits were successfully performed. These volunteers provided informed consent and were compensated at the applicable minimum wage. Participants in the sensitivity evaluation were informed of the persuasive nature of the stimuli, and no personally identifiable information was collected. This study is exempt from IRB review under 45 CFR 46.104(d)(2)(i). We use the PVP dataset in accordance with its CC-BY-4.0 license and use all models and libraries under their respective licenses.

Visual persuasiveness evaluation may support beneficial applications but could also be misused to facilitate misleading or manipulative content. Our framework is intended for analyzing and auditing persuasive reasoning rather than optimizing persuasive content.

REPRODUCIBILITY STATEMENT

The experimental settings and implementation details required to reproduce the main results of this work are described in the paper. Specifically, the experimental setup is provided in Section 3.3, with additional implementation details in Appendix D. We also provide the code and dataset used in our experiments through anonymous links to facilitate reproduction and verification of the results. Together, these materials provide the information needed to reproduce the experimental procedures underlying the main results and conclusions of this work.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan

Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Aysan Aghazadeh and Adriana Kovashka. Cap: Evaluation of persuasive and creative image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16970–16980, 2025.

Jyoti Aneja, Michael Harrison, Neel Joshi, Tyler LaBonte, John Langford, Eduardo Salinas, and Rachel Ward. Phi-4-vision-reasoning technical report. *arXiv:2511.19663*, 2026.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 283–294, 2023.

Jonathan Baron. *Rationality and intelligence*. Cambridge University Press, 2005.

J Anthony Blair. The possibility and actuality of visual arguments. In *Groundwork in the theory of Argumentation: Selected papers of J. Anthony Blair*, pp. 205–223. Springer, 2011.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. Towards cross-tokenizer distillation: the universal logit distillation loss for llms. *arXiv preprint arXiv:2402.12030*, 2024.

Daniel Chandler and Rod Munday. *A dictionary of media and communication*. OUP Oxford, 2011.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4443–4458, 2020.

Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation, 2023. URL https://arxiv.org/abs/2310.19444.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 14852–14882, 2023.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4198–4205, 2020.

Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu. Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 216–223, 2014.

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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19209–19237, 2025.

Jens E Kjeldsen. The study of visual and multimodal argumentation. *Argumentation*, 29(2):115–132, 2015.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pp. 159–174, 1977.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

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/.

Zhexiong Liu, Meiqi Guo, Yue Dai, and Diane Litman. Imagearg: A multi-modal tweet dataset for image persuasiveness mining, 2022. URL https://arxiv.org/abs/2209.06416.

Zhexiong Liu, Mohamed Elaraby, Yang Zhong, and Diane Litman. Overview of imagearg-2023: The first shared task in multimodal argument mining, 2023. URL https://arxiv.org/abs/2310.12172.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1773–1781, 2023.

Edward F McQuarrie and David Glen Mick. Visual rhetoric in advertising: Text-interpretive, experimental, and reader-response analyses. *Journal of consumer research*, 26(1):37–54, 1999.

David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, 9(3):353–383, 1977.

Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8779–8788, 2018.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15012–15032, 2024.

Barbara J Phillips and Edward F McQuarrie. Beyond visual metaphor: A new typology of visual rhetoric in advertising. *Marketing theory*, 4(1-2):113–136, 2004.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: international conference for high performance computing, networking, storage and analysis*, pp. 1–16. IEEE, 2020.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3505–3506, 2020.

Leonard Salewski, A Sophia Koepke, Hendrik PA Lensch, and Zeynep Akata. Clevr-x: A visual reasoning dataset for natural language explanations. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pp. 69–88. Springer, 2020.

Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2591–2600, 2019.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024a.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7059–7073, 2023.

Keith E Stanovich and Richard F West. Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of educational psychology*, 89(2):342, 1997.

Keith E Stanovich, Richard F West, and Maggie E Toplak. Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, 22(4):259–264, 2013.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pp. 613–624. International World Wide Web Conferences Steering Committee, April 2016. doi: 10.1145/2872427.2883081. URL http://dx.doi.org/10.1145/2872427.2883081.

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.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

A Dataset Reconstruction Details

We describe the full annotator-aware binary voting procedure used to reconstruct the PVP dataset as a binary classification task.

A.1 Annotator-aware Binary Voting.

The per-annotator mean persuasiveness scores range from 0.02 to 10, with 22.2% of annotators having a mean score below 3 and 12.1% having a mean score above 7. This indicates substantial variation in individual scoring tendencies, with some annotators assigning systematically lower scores and others assigning systematically higher scores. To account for these annotator-specific scoring scales, we convert each annotator's raw score into a binary vote relative to that annotator's own score distribution, rather than averaging raw scores across annotators.

For each annotator a, we compute the first and third quartiles of that annotator's 0–10 scores. Let $r_i^{(a)}$ be annotator a's score for image–message pair i, and let $Q_1^{(a)}$ and $Q_3^{(a)}$ denote the first and third

quartiles of annotator a's scores. We define the annotator-level vote as

$$v_i^{(a)} = \begin{cases} 1 & \text{if } r_i^{(a)} > Q_3^{(a)}, \\ 0 & \text{if } r_i^{(a)} < Q_1^{(a)}. \end{cases}$$

Scores falling between $Q_1^{(a)}$ and $Q_3^{(a)}$ are treated as ambiguous and do not contribute a vote.

A.2 Consensus-based Label Aggregation.

Image–message pairs for which at least 75% of the collected binary votes are persuasive receive a persuasive label (1), while pairs for which at least 75% are unpersuasive receive an unpersuasive label (0). Pairs that do not meet either threshold are discarded. This procedure retains examples with stronger annotator consensus while allowing for individual differences in the use of the original 0–10 rating scale.

A.3 Inter-Annotator Agreement.

We examine inter-annotator agreement at each stage of the reconstruction procedure. Krippendorff's α is 0.151 on the original 0–10 ratings, reflecting the substantial variation in persuasiveness judgments also reported in the original PVP study. After converting the ratings into annotator-specific binary votes, Krippendorff's α increases to 0.292. On the final set of pairs retained by the 75% consensus criterion, Krippendorff's α, computed over the annotator-level binary votes, further increases to 0.500, with a mean vote agreement of 0.875. This progression indicates that the reconstruction procedure retains examples with stronger agreement while preserving the annotator-aware formulation of the original ratings.

A.4 Dataset Split.

We split the retained pairs by message, ensuring that no message appears in both splits, while maintaining persuasiveness label balance between splits. This message-disjoint split enables evaluation of out-of-message generalization. The final training split contains 820 image–message pairs covering 309 messages, and the test split contains 209 image–message pairs covering 79 messages.

B Prompt Templates for Reasoning Data Collection

This appendix provides the prompt templates used to collect visual persuasion reasoning. Each prompt is conditioned on an input image, an intended message, and a target binary persuasiveness label. The placeholder {message} is replaced with the intended message associated with the image. The token <image> denotes the image input to the vision-language teacher model.

B.1 One-sided Holistic Reasoning

B.1.1 Short rationale: persuasive case

```
<image>
Message: {message}

This image conveys the intended message persuasively.
In 1-2 sentences, explain why this image is persuasive.

Output exactly in this format:
<reasoning>
[Your explanation here]
</reasoning>
```

B.1.2 SHORT RATIONALE: UNPERSUASIVE CASE

```
<image>
Message: {message}

This image fails to convey the intended message persuasively.
In 1–2 sentences, explain why this image fails to be persuasive.

Output exactly in this format:
<reasoning>
[Your explanation here]
</reasoning>
```

B.1.3 LONG RATIONALE: PERSUASIVE CASE

```
<image>
Message: {message}

This image conveys the intended message persuasively.
In 3–4 sentences, explain why this image is persuasive.

Output exactly in this format:
<reasoning>
[Your explanation here]
</reasoning>
```

B.1.4 LONG RATIONALE: UNPERSUASIVE CASE

```
<image>
Message: {message}

This image fails to convey the intended message persuasively.
In 3–4 sentences, explain why this image fails to be persuasive.

Output exactly in this format:
<reasoning>
[Your explanation here]
</reasoning>
```

B.2 IMAGE-DESCRIPTION-GUIDED REASONING

B.2.1 IMAGE DEPICTION RATIONALE: PERSUASIVE CASE

```
<image>
Message: {message}

This image conveys the intended message persuasively.
In 1–2 sentences, briefly describe what the image depicts.
Then in 1–2 sentences, explain why this image is persuasive.

Output exactly in this format:
<reasoning>
Image description: [1–2 sentences]
```

```
Persuasiveness: [1-2 sentences]
</reasoning>
```

B.2.2 IMAGE DEPICTION RATIONALE: UNPERSUASIVE CASE

```
<image>
Message: {message}

This image fails to convey the intended message persuasively.
In 1-2 sentences, briefly describe what the image depicts.
Then in 1-2 sentences, explain why this image fails to be persuasive.

Output exactly in this format:
<reasoning>
Image description: [1-2 sentences]
Persuasiveness: [1-2 sentences]
</reasoning>
```

B.3 ELEMENT-LEVEL REASONING

B.3.1 SINGLE VISUAL ELEMENT RATIONALE: PERSUASIVE CASE

```
<image>
Message: {message}

This image conveys the intended message persuasively.
Identify one specific visual element that most strongly supports its
persuasiveness.
Then explain why that element makes the image persuasive.

Output exactly in this format:
<reasoning>
[visual element]: [reason]
</reasoning>
```

B.3.2 SINGLE VISUAL ELEMENT RATIONALE: UNPERSUASIVE CASE

```
<image>
Message: {message}

This image fails to convey the intended message persuasively.
Identify one specific visual element that most strongly undermines its
     persuasiveness.
Then explain why that element makes the image unpersuasive.

Output exactly in this format:
<reasoning>
[visual element]: [reason]
</reasoning>
```

B.3.3 MULTIPLE VISUAL ELEMENTS RATIONALE: PERSUASIVE CASE

```
<image>
Message: {message}

This image conveys the intended message persuasively.
Examine the visual elements and identify three specific elements that
    strengthen its persuasiveness.

Output exactly in this format:
<reasoning>
Rationale 1:
[visual element]: [reason]
Rationale 2:
[visual element]: [reason]
Rationale 3:
[visual element]: [reason]
</reasoning>
```

B.3.4 MULTIPLE VISUAL ELEMENTS RATIONALE: UNPERSUASIVE CASE

```
<image>
Message: {message}

This image fails to convey the intended message persuasively.
Examine the visual elements and identify three specific elements that
    undermine
its persuasiveness.

Output exactly in this format:
<reasoning>
Rationale 1:
[visual element]: [reason]
Rationale 2:
[visual element]: [reason]
Rationale 3:
[visual element]: [reason]
</reasoning>
```

B.4 SUPPORT–WEAKNESS REASONING

B.4.1 BALANCED RATIONALE: PERSUASIVE CASE

```
<image>
Message: {message}

This image conveys the intended message persuasively.
In one sentence, explain what makes this image persuasive.
Then in one sentence, reflect on any aspect that might still weaken
    the message.
Finally, in one to two sentences, summarize overall why this image is
    persuasive
based on the positive and negative aspects.

Output exactly in this format:
<reasoning>
```

```
Positive: [one sentence]
Negative: [one sentence]
Overall reasoning: [one-two sentences]
</reasoning>
```

B.4.2 BALANCED RATIONALE: UNPERSUASIVE CASE

```
<image>
Message: {message}

This image fails to convey the intended message persuasively.
In one sentence, explain why this image fails to be persuasive.
Then in one sentence, reflect on any aspect that might still support
    the message
. Finally, in one to two sentences, summarize overall why this image
    is
unpersuasive based on the negative and positive aspects.

Output exactly in this format:
<reasoning>
Negative: [one sentence]
Positive: [one sentence]
Overall reasoning: [one-two sentences]
</reasoning>
```

B.5 ELEMENT-LEVEL SUPPORT–WEAKNESS REASONING

B.5.1 ELEMENT-LEVEL SUPPORT–WEAKNESS RATIONALE: PERSUASIVE CASE

```
<image>
Message: {message}

This image conveys the intended message persuasively.
Examine the visual elements and identify one specific element that
    strengthens
its persuasiveness. Then, critically reflect on whether any element
    might still
weaken the message despite its overall effectiveness. Finally, in one
    to two
sentences, summarize overall why this image is persuasive based on the
     positive
and negative aspects.

Output exactly in this format:
<reasoning>
Positive rationale:
[visual element]: [reason]
Negative rationale:
[visual element]: [reason]
Overall reasoning: [one-two sentences]
</reasoning>
```

B.5.2 ELEMENT-LEVEL SUPPORT–WEAKNESS RATIONALE: UNPERSUASIVE CASE

```
<image>
Message: {message}

This image fails to convey the intended message persuasively.
Examine the visual elements and identify one specific element that
    undermines
its persuasiveness. Then, critically reflect on whether any element
    might still
support the message despite its overall ineffectiveness. Finally, in
    one to two
sentences, summarize overall why this image is unpersuasive based on
    the
negative and positive aspects.

Output exactly in this format:
<reasoning>
Negative rationale:
[visual element]: [reason]
Positive rationale:
[visual element]: [reason]
Overall reasoning: [one-two sentences]
</reasoning>
```

C PROMPT TEMPLATES FOR TRAINING

We provide the prompt templates used for each baseline and our proposed method. Base and SFT share the same prompt, as do Base-Reasoning and GRPO.

C.1 BASE AND SFT

```
<image>
Message: {message}
Does this image persuasively convey the message?
Respond only with Yes or No.
```

C.2 BASE-REASONING AND GRPO

Turn 1.

```
<image>
Message: {message}
Consider whether this image persuasively conveys the message.
```

Turn 2.

```
Based on the image, the message, and your reasoning, determine whether
     the image persuasively conveys the message.
Respond only with Yes or No.
```

C.3 GRPO-JOINT

```
<image>
Message: {message}
An image is persuasive if its visual content makes the message more
    convincing
or more emotionally compelling. Given the image and the message,
    determine
whether the image persuasively conveys the message. First, reason
    about the
image's persuasiveness. Then answer Yes or No.

Output exactly in this format:
<reasoning>
[Your reasoning here]
</reasoning>
<answer>
[Yes or No]
</answer>
```

C.4 REASONING-SFT

Turn 1.

```
<image>
Message: {message}
Consider whether this image persuasively conveys the message.
```

Turn 2.

```
Based on the image, the message, and your reasoning, determine whether
     the image persuasively conveys the message.
```

D IMPLEMENTATION DETAILS

D.1 SUPERVISED FINE-TUNING

We fine-tune four vision–language models, Qwen2.5-VL-7B-Instruct (Team, 2025), Phi-3.5-Vision-Instruct (Abdin et al., 2024), Gemma-3-4B-IT (Team et al., 2025), and LLaVA-1.6-Vicuna-7B (Liu et al., 2024), using LoRA (Low-Rank Adaptation) (Hu et al., 2021). Model-specific implementation details are reported in Table 2. For all models, LoRA adapters ($r = 32$, $\alpha = 64$, dropout 0.01) are applied to all linear layers in the language model except *lm_head* and *embed_tokens*, while the base language-model, vision encoder, project and merger weights are frozen. We optimize with AdamW (Loshchilov & Hutter, 2019) using a cosine learning-rate schedule, a peak LoRA learning rate of 1×10^{-4}, 3% warmup, and no weight decay. All runs use an effective batch size of 32, bf16 precision, gradient checkpointing, and DeepSpeed ZeRO-2 (Rasley et al., 2020; Rajbhandari et al., 2020) on a single GPU. Qwen2.5-VL-7B-Instruct and Gemma-3-4B-IT are fine-tuned on an A100 GPU, while Phi-3.5-vision-instruct and LLaVA-1.6-7B are fine-tuned on an RTX A6000 GPU.

For model selection, we reserve 10% of the training set as a validation set, train each model for up to 10 epochs, and select the epoch with the best validation performance. We then retrain each model on the full training set for the selected number of epochs and report performance on the held-out test set.

D.2 Group Relative Policy Optimization

We train the same two student models using Group Relative Policy Optimization (GRPO) (Shao et al., 2024b) with LoRA adapters ($r = 32$, $\alpha = 64$, dropout 0.01) applied to all linear layers in the language model except *lm_head* and *embed_tokens*. All non-LoRA components, including the vision encoder, projector/merger, and base language-model weights, are frozen.

We implement two variants that differ in where the reward is applied. GRPO uses a two-turn structure: the model generates a rationale in Turn 1 and a single-token answer (`Yes` or `No`) in Turn 2, with the gradient applied only to the Turn-2 tokens. GRPO-Joint uses a single-turn structure: the model generates a rationale and answer as one completion in the format `<reasoning>...</reasoning><answer>...</answer>`, with the gradient applied over the full rationale-and-answer sequence.

Both variants use two binary reward functions. The *correctness reward* assigns 1.0 if the answer matches the ground-truth persuasiveness label and 0.0 otherwise. The *format reward* differs between variants: for GRPO, it assigns 1.0 if the Turn-2 output is exactly `Yes` or `No`; for GRPO-Joint, it assigns 1.0 if the single-turn completion follows the full `<reasoning>...</reasoning><answer>...</answer>` structure. The two rewards are combined with weights 1.0 and 0.1 for GRPO, and 1.0 and 0.25 for GRPO-Joint, for correctness and format, respectively. Model-specific implementation details are reported in Table 3. The remaining settings are shared across both variants and both student models: we optimize with AdamW using a cosine learning-rate schedule, a peak LoRA learning rate of 5×10^{-6}, KL penalty $\beta = 0.01$, weight decay 0.1, and 3% warmup. All runs use $G = 4$ rollouts per prompt, generation temperature 1.2, bf16 precision, gradient checkpointing, and DeepSpeed ZeRO-2 on a single GPU.

For model selection, we reserve 10% of the training set as a validation set, train each model for up to 3 epochs, and select the epoch with the best validation performance. We then retrain each model on the full training set for the selected number of epochs and report performance on the held-out test set. Both Qwen2.5-VL-7B-Instruct and Phi-3.5-vision-instruct are fine-tuned on an A100 GPU.

E Additional Classification Results

E.1 Classification Results Across Random Seeds

Table 4 reports the classification results with standard deviations over five random seeds. Reasoning-SFT achieves the highest balanced accuracy on Qwen, Phi, and Gemma, while remaining competitive with label-only SFT on LLaVA. The same trend is reflected in F1, where Reasoning-SFT performs best on Qwen and Gemma and remains close to the strongest baseline on Phi and LLaVA. Overall, the standard deviations indicate that the gains of Reasoning-SFT are generally stable across seeds, although LLaVA exhibits larger variability than the other student models.

E.2 Prompt-Type Ablation

Table 5 reports the performance of student models fine-tuned on rationales from each individual prompt type, compared with the merged Reasoning-SFT setting. We highlight two findings.

Table 2: Model-specific implementation settings for supervised fine-tuning. Shared settings, including the LoRA configuration, optimizer, learning-rate schedule, effective batch size, and precision, are described in the text.

	Qwen2.5-VL-7B-Instruct	Phi-3.5-vision-instruct
Per-device batch $\times$ grad. accum.	8×4	4×8
Image preprocessing	Patch-grid area $\in [1024, 1280] \cdot 28^2$ px	16 crops

Table 3: Model-specific implementation settings for GRPO and GRPO-Joint. Shared settings are described in the text.

	Qwen2.5-VL-7B-Instruct	Phi-3.5-vision-instruct
Per-device batch × grad. accum.	4×4	1×32
Image preprocessing	Patch-grid area $\in [1024, 1280] \cdot 28^2$ px	16 crops

Table 4: Persuasiveness classification performance (%) across five random seeds. Results are reported as mean ± standard deviation. **Bold** indicates the best BA and F1 within each student model.

Student	Setup	BA	F1	P	R
Qwen	Base	63.2 ± 1.3	68.7 ± 1.5	59.6 ± 0.8	81.3 ± 3.2
	Base-Reasoning	61.9 ± 1.9	68.9 ± 2.0	57.9 ± 2.4	85.0 ± 4.4
	SFT	74.6 ± 1.6	75.0 ± 1.7	75.0 ± 2.3	75.3 ± 1.5
	GRPO-Ans	70.8 ± 2.1	71.3 ± 2.7	70.0 ± 2.3	72.8 ± 4.6
	GRPO	69.2 ± 2.8	73.2 ± 2.7	64.4 ± 2.2	85.0 ± 4.8
	Reasoning-SFT	$\mathbf{75.1 \pm 1.8}$	$\mathbf{75.1 \pm 1.9}$	75.7 ± 2.1	74.7 ± 1.6
Phi	Base	65.2 ± 1.4	69.9 ± 1.4	61.4 ± 1.9	81.1 ± 2.4
	Base-Reasoning	56.9 ± 2.9	64.6 ± 4.3	54.7 ± 3.1	77.1 ± 4.6
	SFT	73.7 ± 1.8	$\mathbf{75.6 \pm 0.5}$	70.5 ± 3.0	81.7 ± 3.7
	GRPO-Ans	67.5 ± 1.2	66.6 ± 2.4	74.3 ± 2.0	60.5 ± 3.5
	GRPO	55.3 ± 2.9	67.6 ± 1.8	52.7 ± 2.9	94.2 ± 3.0
	Reasoning-SFT	$\mathbf{74.8 \pm 1.3}$	74.2 ± 2.5	76.3 ± 1.8	72.2 ± 4.9
Gemma	Base	61.7 ± 2.0	69.9 ± 1.7	57.4 ± 1.4	89.6 ± 3.0
	Base-Reasoning	55.2 ± 1.8	63.9 ± 1.3	53.4 ± 1.5	79.6 ± 1.4
	SFT	72.1 ± 3.8	71.8 ± 4.4	72.2 ± 4.7	71.7 ± 6.4
	GRPO-Ans	71.5 ± 3.1	73.4 ± 3.1	68.5 ± 3.0	79.4 ± 6.7
	GRPO	66.7 ± 2.5	72.7 ± 1.7	61.5 ± 2.0	89.0 ± 3.7
	Reasoning-SFT	$\mathbf{75.7 \pm 2.7}$	$\mathbf{75.3 \pm 2.3}$	76.7 ± 4.6	74.3 ± 4.8
LLaVA	Base	50.6 ± 0.5	66.7 ± 1.4	50.1 ± 1.5	99.8 ± 0.4
	Base-Reasoning	52.0 ± 2.4	66.1 ± 2.2	50.9 ± 2.0	94.3 ± 2.6
	SFT	$\mathbf{72.9 \pm 5.9}$	$\mathbf{74.6 \pm 5.2}$	70.3 ± 7.1	79.9 ± 6.8
	GRPO-Ans	53.2 ± 6.0	67.2 ± 3.9	51.7 ± 4.6	96.5 ± 1.6
	GRPO	56.1 ± 3.8	68.5 ± 2.6	53.2 ± 3.1	96.3 ± 2.4
	Reasoning-SFT	72.5 ± 4.4	72.5 ± 3.9	72.7 ± 6.2	72.6 ± 5.5

E.2.1 Combining reasoning perspectives improves balanced accuracy

The Reasoning-SFT model achieves the highest balanced accuracy for both student models, and the performance gain is not explained by any single perspective alone. We also examined whether performance differences are explained by rationale length (see Appendix E.3): the Pearson correlation between average rationale length and balanced accuracy is low for both students, confirming that content and diversity matter more than length.

E.2.2 The strongest single-prompt variant is inconsistent across student models.

Among single-prompt variants, Long performs best on Qwen2.5-VL-7B-Instruct in both balanced accuracy and F1 score. On Phi-3.5-vision-instruct, however, Short gives the highest balanced accuracy while Long gives the highest F1 score. This inconsistency indicates that no single prompt provides a reliable substitute for Reasoning-SFT across both student models and metrics.

E.3 Rationale Length and Performance Across Prompt Types

Average rationale length does not consistently account for the performance differences across rationale sources. As shown in Table 6, the strongest results are not obtained by the longest rationales:

Table 5: Effect of prompt type on rationale-based fine-tuning. Each row reports student model performance after fine-tuning on rationales generated by the corresponding prompt. In the Axis column, S and C denote Support-Focused and Counter-Aware, respectively, while G and L denote Global and Local, respectively. **Bold** indicates the best score per model and metric.

		Qwen2.5-VL-7B-Instruct		Phi-3.5-Vision-Instruct	
Axis	**Prompt type**	**Balanced Acc.**	**F1 Score**	**Balanced Acc.**	**F1 Score**
S×G	Short	0.707	0.707	0.700	0.687
	Long	0.741	0.741	0.684	**0.749**
	Description	0.670	0.667	0.685	0.683
S×L	Visual Element	0.662	0.660	0.600	0.567
	Visual Elements	0.698	0.698	0.622	0.633
C×G	Balanced	0.667	0.664	0.666	0.718
C×L	Visual Elements	0.670	0.667	0.694	0.698
	Reasoning-SFT	**0.766**	**0.766**	**0.746**	0.746

Table 6: Performance of different rationale sources with their average rationale lengths. Across both student models, the strongest results are not obtained by the longest rationales, suggesting that rationale length alone does not explain the performance differences. In the Axis column, S and C denote Support-Focused and Counter-Aware, respectively, while G and L denote Global and Local, respectively.

		Qwen2.5-VL-7B-Instruct			Phi-3.5-Vision-Instruct		
Axis	**Rationale source**	**Avg. tokens**	**Bal. Acc.**	**F1**	**Avg. tokens**	**Bal. Acc.**	**F1**
S×G	Short	65.0	0.707	0.707	77.1	0.700	0.687
	Long	91.2	0.741	0.741	130.7	0.684	0.749
	Description	113.0	0.670	0.667	127.8	0.685	0.683
S×L	Visual Element	51.6	0.662	0.660	44.7	0.600	0.567
	Visual Elements	164.2	0.698	0.698	139.1	0.622	0.633
C×G	Balanced	102.8	0.667	0.664	124.7	0.666	0.718
C×L	Visual Elements	125.7	0.670	0.667	131.3	0.694	0.698
	Reasoning-SFT	101.9	0.766	0.766	110.8	0.746	0.746
Pearson r with avg. tokens		–	0.236	0.238	–	0.305	0.588

across both student models, Reasoning-SFT achieves the best balanced accuracy despite not having the longest rationales. The Pearson correlation between average token length and balanced accuracy is 0.236 for Qwen2.5-VL-7B-Instruct and 0.305 for Phi-3.5-vision-instruct, confirming that rationale length alone does not explain the observed performance differences.

F Theoretical and Empirical Analysis of Reasoning-Perspective Diversity

F.1 Coverage-Based Motivation for Dual-Axis Rationale Supervision

The preceding experiments suggest that training on rationales from multiple prompt types is more effective than relying on a single source. We provide a simple coverage-based motivation for this effect: even when all rationales support the same target label, different prompt types may expose complementary valid reasoning paths.

For an image–message input x with target label $y^\star(x)$, let $\mathcal{C}(x)$ denote the set of valid rationales that support $y^\star(x)$. Different prompt types are therefore treated as alternative valid reasoning paths rather

than contradictory supervision. If a selected rationale set covers this space well, low training loss on the selected rationales also constrains the loss on nearby valid reasoning paths. Let $S(x) \subseteq \mathcal{C}(x)$ be the selected training rationale set, and let $\psi : \mathcal{C}(x) \to \mathbb{R}^d$ be a fixed embedding map. We define its coverage radius as

$$\rho(S; x) = \sup_{r \in \mathcal{C}(x)} \min_{s \in S(x)} \|\psi(r) - \psi(s)\|_2.$$

Let θ denote the model parameters, and let $\ell(\theta; x, r)$ denote the training loss when rationale r, together with the target label, is used as supervision for input x. The following theorem formalizes the coverage intuition under an idealized Lipschitz assumption.

Theorem 1 (Rationale coverage bound)**.** *Assume that $\ell(\theta; x, r)$ is L-Lipschitz in the rationale embedding for some constant $L > 0$, i.e.,*

$$|\ell(\theta; x, r) - \ell(\theta; x, r')| \leq L\|\psi(r) - \psi(r')\|_2 \quad \textit{for all } r, r' \in \mathcal{C}(x).$$

Then, for any selected rationale set $S(x) \subseteq \mathcal{C}(x)$,

$$\sup_{r \in \mathcal{C}(x)} \ell(\theta; x, r) \leq \max_{s \in S(x)} \ell(\theta; x, s) + L\rho(S; x).$$

Theorem 1 provides a formal perspective on why complementary rationale supervision can help: with comparable empirical fit, broader coverage of valid reasoning paths yields a tighter worst-case loss bound. Our dual-axis design follows this principle by varying evidence polarity and visual granularity.

We further analyze why rationale supervision from complementary reasoning perspectives can be useful. Building on the coverage view above, we first provide the proof of Theorem 1 and empirically test whether the proposed dual-axis design provides broader coverage under matched rationale budgets. We then consider two complementary perspectives—spectral conditioning and redundancy reduction—and report additional diagnostics of the resulting rationale-source family.

Valid rationale space. For each image–message input x, let $y^\star(x) \in \mathcal{Y}$ denote the target persuasiveness label, and let $\mathcal{C}(x)$ denote the latent set of valid rationales supporting $y^\star(x)$. All rationales considered for the same input are assumed to support the same target label. Thus, the relevant diversity arises from alternative valid reasoning paths rather than contradictory supervision.

Rationale-source family. For each input x, we consider a family of rationale-generation sources

$$\mathcal{F}(x) = \{R_1(x), \ldots, R_K(x)\},$$

where each $R_k(x) \subseteq \mathcal{C}(x)$ corresponds to a prompt type or reasoning perspective. Let

$$\mathcal{U}(x) = \bigcup_{k=1}^{K} R_k(x)$$

denote the generated candidate pool. For a fixed rationale budget B, we compare source compositions while holding the number of selected rationales fixed.

Rationale embedding. Let

$$\psi(r) \in \mathbb{R}^d$$

denote a fixed embedding of rationale r. Since embedding-based diversity measures may depend on the representation space, we evaluate the empirical proxies under multiple embedding backends.

F.2 Coverage Diversity

Coverage measures how well a selected rationale set represents alternative valid reasoning paths. For a selected set $S(x) \subseteq \mathcal{C}(x)$, the coverage radius introduced in the main text is

$$\rho(S; x) = \sup_{r \in \mathcal{C}(x)} \min_{s \in S(x)} \|\psi(r) - \psi(s)\|_2.$$

A smaller radius means that every valid rationale is closer to at least one selected rationale.

Table 7: Matched-budget coverage of same-axis and dual-axis rationale subsets. Coverage is the average nearest-subset cosine distance to the teacher-rationale candidate pool; lower is better. Relative reduction is measured against the same-axis subset.

Generator	k	Same-axis	Dual-axis	Reduction
Qwen	2	0.137	0.127	7.4%
Qwen	3	0.134	0.117	13.1%
Phi	2	0.158	0.134	15.0%
Phi	3	0.154	0.121	21.3%

Proof of Theorem 1. Fix any $r \in \mathcal{C}(x)$. By the definition of $\rho(S; x)$, there exists $s_r \in S(x)$ such that

$$\|\psi(r) - \psi(s_r)\|_2 \leq \rho(S; x).$$

Under the Lipschitz assumption,

$$\ell(\theta; x, r) \leq \ell(\theta; x, s_r) + L\|\psi(r) - \psi(s_r)\|_2 \leq \max_{s \in S(x)} \ell(\theta; x, s) + L\rho(S; x).$$

Taking the supremum over $r \in \mathcal{C}(x)$ yields

$$\sup_{r \in \mathcal{C}(x)} \ell(\theta; x, r) \leq \max_{s \in S(x)} \ell(\theta; x, s) + L\rho(S; x),$$

which proves the claim.

Empirical coverage proxies. The full valid-rationale space $\mathcal{C}(x)$ is unobserved, so we approximate it using the generated candidate pool $\mathcal{U}(x)$. For a selected rationale set $S(x)$, we define

$$\rho_{\text{avg}}(S; x) = \frac{1}{|\mathcal{U}(x)|} \sum_{u \in \mathcal{U}(x)} \min_{s \in S(x)} d(\psi(u), \psi(s)), \tag{1}$$

$$\rho_{\max}(S; x) = \max_{u \in \mathcal{U}(x)} \min_{s \in S(x)} d(\psi(u), \psi(s)), \tag{2}$$

where d denotes the embedding distance used in each analysis. Smaller values indicate better coverage of the generated rationale space.

For normalized embeddings, cosine distance is monotonic in Euclidean distance, since

$$\|\psi(u) - \psi(s)\|_2^2 = 2\big(1 - \cos(\psi(u), \psi(s))\big).$$

Thus, the cosine-distance proxy used below preserves the ordering induced by Euclidean distance in the coverage bound.

Matched-budget dual-axis coverage. We first test the prediction most directly relevant to our rationale design: whether combining the two axes provides broader coverage than selecting rationales restricted to a single axis. For each input, we compare same-axis and dual-axis source subsets under the same rationale budget k, using average nearest-subset cosine distance to the full teacher-rationale pool as the coverage measure. For each k, we average the coverage distance over all eligible same-axis and dual-axis source subsets and then average across inputs.

Dual-axis subsets achieve lower coverage distance for both teacher families and both budgets, reducing the distance by 7.4–13.1% for Qwen and 15.0–21.3% for Phi. All differences are statistically significant under 10,000 paired bootstrap resamples. At $k = 4$, dual-axis subsets also outperform all non-dual subsets, with relative reductions of 7.5% for Qwen and 10.2% for Phi.

We further examine whether this coverage measure is related to downstream performance. Across single-prompt rationale sources, lower coverage distance is strongly associated with higher balanced accuracy after fine-tuning (Spearman $\rho = 0.955$ for Qwen and $\rho = 0.786$ for Phi), and the relationship remains robust to rationale length. Together, these results support the coverage hypothesis underlying our dual-axis design: combining evidence polarity and visual granularity exposes complementary reasoning paths that cover the generated rationale space more effectively.

Table 8: Random-baseline coverage as a function of budget B, averaged over 10 draws per input. Coverage is computed against the common candidate pool of all available Phi and Qwen rationale sources for the same input. Smaller values indicate better coverage.

Embedding	Split	$\rho_{\text{avg}} \downarrow$ for $\texttt{rand}(B)$					$\rho_{\max} \downarrow$ for $\texttt{rand}(B)$				
		B=1	B=2	B=3	B=4	B=5	B=1	B=2	B=3	B=4	B=5
BGE-M3	Phi	0.0976	0.0730	0.0605	0.0512	0.0436	0.1669	0.1416	0.1294	0.1209	0.1144
BGE-M3	Qwen	0.0934	0.0727	0.0605	0.0515	0.0441	0.1657	0.1485	0.1402	0.1345	0.1299
E5	Phi	0.0616	0.0460	0.0378	0.0318	0.0269	0.1083	0.0924	0.0849	0.0803	0.0769
E5	Qwen	0.0596	0.0451	0.0373	0.0314	0.0267	0.1041	0.0907	0.0843	0.0801	0.0767
MXBAI	Phi	0.1138	0.0839	0.0692	0.0587	0.0503	0.2119	0.1769	0.1613	0.1515	0.1444
MXBAI	Qwen	0.1101	0.0824	0.0683	0.0584	0.0503	0.2047	0.1727	0.1591	0.1504	0.1444

Coverage scaling with rationale budget. As a complementary sanity check, we examine how coverage changes as the number of sampled rationale sources increases. For each input x, the random baseline uniformly samples B sources from the available rationale family, and coverage is measured against the common candidate pool of all available rationales for that input.

Increasing B consistently reduces both average and worst-case coverage distance across embedding backends and teacher families. This confirms the expected monotonic behavior of the coverage proxy and complements the matched-budget analysis above, which isolates the effect of source composition rather than rationale count alone.

F.3 Spectral Conditioning Diversity

Coverage describes whether selected rationales reach different parts of the rationale space, but does not characterize whether they span multiple independent directions or remain concentrated along a narrow semantic axis. To capture this distinction, let $S(x) = \{r_1, \dots, r_m\}$, define $z_i = \psi(r_i)$, and let

$$\Sigma_{S(x)} = \frac{1}{m}\sum_{i=1}^{m}(z_i - \bar{z})(z_i - \bar{z})^\top, \qquad \bar{z} = \frac{1}{m}\sum_{i=1}^{m} z_i.$$

Proposition 1 (Spectral diversity and ridge-stabilized local estimation)**.** *Assume a local linearization of the fine-tuning objective around a reference parameter θ_0, so that the rationale-conditioned prediction is approximated by*

$$f_\theta(x, r) \approx f_{\theta_0}(x, r) + h(x, r)^\top(\theta - \theta_0),$$

where $h(x, r) \in \mathbb{R}^p$ is a local feature or Jacobian vector. Consider the ridge-regularized local objective over $S(x) = \{r_1, \dots, r_m\}$:

$$\hat{\theta} = \arg\min_\theta \sum_{i=1}^{m}\big(y_i - h(x, r_i)^\top\theta\big)^2 + \lambda\|\theta\|_2^2, \qquad \lambda > 0.$$

Let $H \in \mathbb{R}^{m\times p}$ be the design matrix with rows $h(x, r_i)^\top$. If

$$y_i = h(x, r_i)^\top\theta^\star + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i] = 0, \qquad \mathrm{Var}(\varepsilon_i) = \sigma^2,$$

then the variance component of the ridge estimation error satisfies

$$\sigma^2\mathrm{tr}\big((H^\top H + \lambda I)^{-2}H^\top H\big) \le \sigma^2\mathrm{tr}\big((H^\top H + \lambda I)^{-1}\big).$$

The full mean-squared error satisfies

$$\mathbb{E}\|\hat{\theta} - \theta^\star\|_2^2 \le \sigma^2\mathrm{tr}\big((H^\top H + \lambda I)^{-1}\big) + \lambda^2\|\theta^\star\|_2^2\,\big\|(H^\top H + \lambda I)^{-1}\big\|_2^2.$$

Thus, a rationale set whose local design matrix has a richer and less collapsed spectrum yields a better-conditioned local estimation problem.

Proof. The ridge estimator satisfies

$$\hat{\theta} - \theta^\star = (H^\top H + \lambda I)^{-1}(H^\top \varepsilon - \lambda\theta^\star),$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)^\top$. Using $\mathbb{E}[\varepsilon\varepsilon^\top] = \sigma^2 I_m$, the variance term is

$$\sigma^2 \mathrm{tr}\big((H^\top H + \lambda I)^{-2} H^\top H\big).$$

Since

$$H^\top H \preceq H^\top H + \lambda I,$$

we have

$$(H^\top H + \lambda I)^{-1} H^\top H (H^\top H + \lambda I)^{-1} \preceq (H^\top H + \lambda I)^{-1}.$$

Taking traces gives the stated variance bound. The bias term satisfies

$$\lambda^2 \theta^{\star\top} (H^\top H + \lambda I)^{-2} \theta^\star \le \lambda^2 \|\theta^\star\|_2^2 \left\|(H^\top H + \lambda I)^{-1}\right\|_2^2.$$

Combining the variance and bias terms proves the result. □

Remark 1. *The proposition is stated in terms of local Jacobian directions $h(x, r)$, whereas our empirical diagnostics use external rationale embeddings $\psi(r)$. We therefore use embedding covariance only as a gradient-free proxy for diversity in local supervision directions.*

Spectral proxies. We measure the spectrum of $\Sigma_{S(x)}$ using

$$\mathrm{erank}(\Sigma_{S(x)}) = \frac{\mathrm{tr}(\Sigma_{S(x)})}{\|\Sigma_{S(x)}\|_2}, \tag{3}$$

$$D_{\log\det}(S(x)) = \log\det(I + \alpha\Sigma_{S(x)}), \qquad \alpha > 0, \tag{4}$$

where larger values indicate greater multi-directional or volumetric spread. We additionally report

$$\mathrm{PR}(\Sigma_{S(x)}) = \frac{\big(\mathrm{tr}(\Sigma_{S(x)})\big)^2}{\mathrm{tr}(\Sigma_{S(x)}^2)}, \tag{5}$$

$$A(S(x)) = \frac{\lambda_{\max}(\Sigma_{S(x)})}{\mathrm{tr}(\Sigma_{S(x)})}, \tag{6}$$

where smaller A indicates less directional concentration.

F.4 Redundancy Reduction

A rationale family may provide broad coverage while still containing substantial redundancy, for example when multiple rationales are paraphrases of the same reasoning pattern. We therefore consider a complementary variance-reduction view.

Proposition 2 (Variance reduction from low-redundancy supervision)**.** *Suppose each rationale provides a noisy supervision signal*

$$z_r(x) = f^\star(x) + \varepsilon_r(x), \qquad \mathbb{E}[\varepsilon_r(x) \mid x] = 0, \qquad \mathrm{Var}(\varepsilon_r(x) \mid x) = \sigma^2(x).$$

Let $S(x) = \{r_1, \ldots, r_m\}$ and define

$$\bar{z}(x) = \frac{1}{m}\sum_{i=1}^{m} z_{r_i}(x).$$

If the average pairwise noise correlation is

$$\bar{\rho}(x) = \frac{2}{m(m-1)} \sum_{1 \le i < j \le m} \mathrm{Corr}\left(\varepsilon_{r_i}(x), \varepsilon_{r_j}(x) \mid x\right),$$

then

$$\mathrm{Var}(\bar{z}(x) \mid x) = \frac{\sigma^2(x)}{m}\left(1 + (m-1)\bar{\rho}(x)\right).$$

Therefore, under a matched rationale budget, a less redundant supervision set can yield lower effective supervision variance than a highly correlated set.

Table 9: Rationale-source diversity proxies used in the empirical analysis.

Proxy	Meaning	Preferred direction
$\mathrm{Sim}_{\mathrm{avg}}$	average pairwise cosine similarity	smaller = less redundant
$\mathrm{cos_dist}_{\mathrm{avg}}$	average pairwise cosine distance	larger = less redundant
erank	effective rank of covariance	larger = more multi-directional
logdet	volumetric spread of covariance	larger = more volumetric
anisotropy	directional concentration	smaller = more isotropic
ρ_{avg}	average coverage distance	smaller = better coverage
$\rho_{\max}$	worst-case coverage distance	smaller = better coverage
near-dup	near-duplicate rate	smaller = less redundant

Proof. Using the variance of an average,

$$\mathrm{Var}(\bar{z}(x) \mid x) = \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}(z_{r_i}(x) \mid x) + \frac{2}{m^2}\sum_{1\le i<j\le m}\mathrm{Cov}\left(z_{r_i}(x), z_{r_j}(x) \mid x\right).$$

Each variance term equals $\sigma^2(x)$, while each covariance term is

$$\mathrm{Corr}\left(\varepsilon_{r_i}(x), \varepsilon_{r_j}(x) \mid x\right)\sigma^2(x).$$

Substituting these terms and simplifying yields

$$\mathrm{Var}(\bar{z}(x) \mid x) = \frac{\sigma^2(x)}{m}\left(1 + (m-1)\bar{\rho}(x)\right).$$

□

Remark 2. *In visual persuasion, rationales can focus on global composition, local visual elements, or supporting and weakening evidence. These rationales may share the same target label while relying on different evidence channels, making reduced redundancy a plausible benefit of complementary supervision.*

Redundancy proxies. For $S(x) = \{r_1, \ldots, r_m\}$, we use

$$D_{\mathrm{pair}}(S(x)) = \frac{2}{m(m-1)}\sum_{1\le i<j\le m}\|\psi(r_i) - \psi(r_j)\|_2^2, \tag{7}$$

$$\mathrm{Sim}_{\mathrm{avg}}(S(x)) = \frac{2}{m(m-1)}\sum_{1\le i<j\le m}\cos(\psi(r_i), \psi(r_j)). \tag{8}$$

Larger D_{pair} and smaller $\mathrm{Sim}_{\mathrm{avg}}$ indicate lower redundancy.

F.5 Empirical Diagnostics of Rationale Sources

We next examine whether the prompt-specific rationale sources exhibit systematic structure beyond the matched-budget coverage results above. We evaluate source distinctness, relationships among diversity proxies, pairwise source structure, and robustness across embedding backends.

Proxy summary. Table 9 summarizes the diversity proxies used in the empirical analysis.

Source distinctness. For each input x, let $r_k(x)$ be the rationale produced by source k, and let

$$e_k(x) = \psi(r_k(x)).$$

To remove input-specific effects, we residualize each source embedding within the input:

$$\tilde{e}_k(x) = e_k(x) - \frac{1}{K_x}\sum_{j=1}^{K_x} e_j(x),$$

Table 10: Global PERMANOVA after input-wise mean removal and restricted within-input permutations. In all cases, $p = 0.005$ under 199 permutations. N denotes the number of residual samples and K the number of co-occurring rationale sources.

Embedding	**Generator**	N	K	F	R^2
E5	Qwen	5670	7	1235.45	0.5669
E5	Phi	5738	7	1216.27	0.5601
BGE-M3	Qwen	5670	7	1089.71	0.5359
BGE-M3	Phi	5738	7	1004.06	0.5125
MXBAI	Qwen	5670	7	644.29	0.4057
MXBAI	Phi	5738	7	560.99	0.3700

where K_x is the number of available sources for input x. We then run PERMANOVA on $\tilde{e}_k(x)$, restricting permutations within each input block.

Prompt-source identity explains a substantial portion of the residual embedding variation after controlling for input identity, indicating that the seven rationale sources are not simply interchangeable paraphrases.

Relationships among diversity proxies. For two image-level proxy functions M_1 and M_2, define

$$\rho(M_1, M_2) = \text{Spearman}\left(\{M_1(S(x))\}_{x\in\mathcal{D}}, \{M_2(S(x))\}_{x\in\mathcal{D}}\right).$$

Under BGE-M3 embeddings,

$$\rho(\text{erank}, \rho_{\text{avg}}) \approx -0.006 \quad \text{on Phi}, \qquad \rho(\text{erank}, \rho_{\text{avg}}) \approx -0.034 \quad \text{on Qwen},$$

while

$$\rho(\text{erank}, \text{Sim}_{\text{avg}}) \approx 0.303 \quad \text{on Phi}, \qquad \rho(\text{erank}, \text{Sim}_{\text{avg}}) \approx 0.166 \quad \text{on Qwen}.$$

These weak-to-moderate relationships indicate that coverage, spectral spread, and redundancy capture related but non-equivalent properties of rationale diversity.

Pairwise source structure. For each source pair (k, ℓ), define

$$C_{k\ell} = \frac{1}{|\mathcal{D}_{k\ell}|} \sum_{x\in\mathcal{D}_{k\ell}} \cos\left(\psi(r_k(x)), \psi(r_\ell(x))\right), \qquad D_{k\ell} = 1 - C_{k\ell}.$$

We also compute the near-duplicate rate

$$N^{(\tau)}_{k\ell} = \frac{1}{|\mathcal{D}_{k\ell}|} \sum_{x\in\mathcal{D}_{k\ell}} \mathbf{1}\left[\cos\left(\psi(r_k(x)), \psi(r_\ell(x))\right) \geq \tau\right],$$

with $\tau = 0.95$. These quantities define the source-pair matrices

$$C = (C_{k\ell})_{k,\ell=1}^{K}, \qquad D = (D_{k\ell})_{k,\ell=1}^{K}, \qquad N^{(\tau)} = (N^{(\tau)}_{k\ell})_{k,\ell=1}^{K}.$$

For example, in the Qwen rationale family, `long` and `short` remain relatively close under multiple embedding backends, whereas `balanced` and `visual_element` are farther apart. This indicates that the source family contains both partially redundant and complementary rationale pairs.

Robustness across embedding backends. For embedding backends a and b and a proxy M, define

$$\rho_{a,b}(M) = \text{Spearman}\left(\{M_a(S(x))\}_{x\in\mathcal{D}}, \{M_b(S(x))\}_{x\in\mathcal{D}}\right).$$

For pairwise source matrices, we compute Spearman correlation over the upper-triangular entries:

$$\rho^{\text{pair}}_{a,b}(C) = \text{Spearman}\left(\{C^{(a)}_{k\ell} : 1 \leq k < \ell \leq K\}, \{C^{(b)}_{k\ell} : 1 \leq k < \ell \leq K\}\right).$$

The pairwise cosine-similarity structures show moderate-to-strong agreement across embedding backends. Per-image spectral proxies are somewhat more backend-sensitive, especially for Qwen, but remain positively correlated. Near-duplicate rates are less stable in some cases because they depend on the fixed threshold $\tau = 0.95$.

Table 11: Embedding-backend robustness of ψ-based proxies using Spearman correlations. Per-image correlations are computed over the common $n = 820$ inputs; pairwise matrix correlations use the upper-triangular entries of the source-pair matrices.

Split	Quantity	ρ(E5,BGE)	ρ(E5,MXBAI)	ρ(BGE,MXBAI)
Qwen	per-image $\mathrm{sim}_{\mathrm{avg}}$	0.602	0.745	0.755
Qwen	per-image logdet	0.603	0.745	0.755
Qwen	per-image erank	0.540	0.583	0.527
Qwen	per-image $\mathrm{anisotropy}$	0.540	0.583	0.527
Qwen	pairwise $\mathrm{cos_mean}$ matrix	0.809	0.955	0.806
Qwen	pairwise near-duplicate matrix	0.506	0.891	0.518
Phi	per-image $\mathrm{sim}_{\mathrm{avg}}$	0.609	0.727	0.813
Phi	per-image logdet	0.609	0.728	0.813
Phi	per-image erank	0.736	0.691	0.769
Phi	per-image $\mathrm{anisotropy}$	0.736	0.691	0.769
Phi	pairwise $\mathrm{cos_mean}$ matrix	0.747	0.892	0.831
Phi	pairwise near-duplicate matrix	0.774	0.890	0.747

Takeaway. Together, these analyses provide complementary support for dual-axis rationale supervision. The coverage bound explains why broader coverage can make a selected rationale set more informative, while the matched-budget experiments show that combining evidence polarity and visual granularity indeed improves coverage of the generated rationale space. The spectral and redundancy analyses further show that prompt sources differ in their directional spread and overlap rather than forming a collection of interchangeable paraphrases. Overall, the theoretical and empirical results support the use of label-consistent, complementary reasoning perspectives as a richer supervision signal for visual persuasion.

G Details on Rationale Faithfulness Evaluation

G.1 Ground Truth Dataset Construction

To establish the evaluation methodology for *groundedness* and *consistency*, we sample 10% of instances from the rationale training data. To establish ground-truth labels, three of the authors independently labeled each instance. The annotations exhibited substantial agreement on *groundedness* (Fleiss' $\kappa = 0.680$, 88.9% complete agreement) and near-perfect agreement on *consistency* (Fleiss' $\kappa = 0.996$, 99.7% complete agreement) (Landis & Koch, 1977). Given this level of agreement, we determine the ground-truth labels via majority vote among the three annotators.

G.2 Comparison of Groundedness Candidates

Evaluating groundedness in the context of visual persuasion is inherently challenging. Unlike factual image captioning, persuasive rationales often mix objective visual elements with subjective interpretations like emotional tone or symbolism. To capture such nuance, we evaluated five distinct evaluation methods against not only the human majority vote but also the individual judgments of the three annotators.

The candidate methods included:

1. **CLIP with text:** Measuring CLIP similarity between the rationale text and the image description provided in the PVP dataset.
2. **CLIP with image:** Measuring CLIP similarity between the rationale text and the corresponding image.
3. **GPT-5 atomic facts:** Decomposing the rationale into discrete atomic facts and prompting GPT-5 to verify each strictly against the image.
4. **GPT-5 atomic facts (calibrated):** The same atomic facts approach, but explicitly calibrated using the validation set to select a threshold for ratio (N_{yes}/N_{total})

5. **GPT-5 prompting:** A direct, zero-shot prompt asking GPT-5 to act as a judge of the rationale's groundedness in the provided image.

Table 12: Comparison of groundedness evaluation methods.

Method	κ (vs Majority)	κ (vs U1)	κ (vs U2)	κ (vs U3)	Bal. Acc.	F1 (Yes)	F1 (No)
CLIP with image (> 0.3)	-0.0016	0.0647	0.0349	-0.0192	0.4983	0.7414	0.1616
CLIP with text (> 0.6)	0.0406	0.0426	0.0316	0.0516	0.5672	0.5211	0.2138
GPT-5 atomic facts	0.1001	0.1424	0.1114	0.1141	**0.6526**	0.6058	**0.2603**
GPT-5 atomic facts (> 0.6)	0.1252	0.1351	0.0977	**0.1647**	0.5563	**0.9186**	0.2051
GPT-5 prompting	**0.1451**	**0.1935**	**0.2105**	0.1294	0.5711	0.9125	0.2326

Evaluating a model's ability to recognize ungrounded rationales, measured by the F1 Score (No), is arguably the most critical aspect of this metric, as it indicates the model's sensitivity to hallucinations. However, an aggressive rejection strategy can artificially inflate this score at the cost of rejecting valid interpretations.

The results demonstrate that traditional vision-language embedding distances (CLIP) perform poorly, yielding near-zero correlation with human judgments and exhibiting extreme imbalances between positive and negative class identification.

Decomposing the rationales into uncalibrated atomic facts achieved the highest $F_1(no)$ Score of 0.2603 and balanced accuracy of 0.6526. However, this method suffered a severe drop in the positive class $F_1(yes)$ Score of 0.6058. This indicates that strict factual decomposition over-penalizes subjective interpretations, generating excessive false negatives by rejecting valid persuasive reasoning. When we explicitly calibrated the atomic facts prompt to account for this subjectivity, the $F_1(yes)$ Score recovered dramatically, but the ability to detect ungrounded rationales degraded, with the $F_1(no)$ Score dropping to 0.2051.

Ultimately, **GPT-5 prompting** provided the most optimal trade-off and robust overall performance. By assessing the rationale holistically rather than breaking it into isolated components, it achieved a strong $F_1(no)$ Score of 0.2326 while maintaining a high $F_1(yes)$ Score of 0.9125. This balanced detection capability resulted in the highest overall agreement with the aggregated majority vote ($\kappa = 0.1451$) and the strongest alignment with individual human annotators (κ of 0.2105 with User 2 and 0.1935 with User 1). Consequently, GPT-5 prompting was selected as the standard groundedness metric for our evaluation pipeline.

G.3 Comparison of Consistency Candidates

To automate the evaluation of consistency and groundedness, we assessed several candidate models and prompting strategies. Our goal was to select the method that achieved the highest alignment with the human annotations. Alignment is primarily measured using Cohen's κ, alongside Balanced Accuracy and F_1 score.

For consistency, the task requires determining whether the persuasiveness label naturally follows from the provided rationale text. To select our evaluator, we compared an LLM-as-a-judge approach using GPT-5, Qwen2.5-VL-32B-Instruct, and Qwen2.5-VL-7B-Instruct against a dedicated Natural Language Inference (NLI) model (`nli-deberta-v3-base`).

Table 13: Comparison of consistency evaluation methods.

Model	Cohen's κ	Balanced Acc.	F1 Score
GPT-5	**0.992**	**0.996**	**0.996**
Qwen3-32B	0.914	0.958	0.958
Qwen3-8B	0.973	0.986	0.986
NLI (DeBERTa)	0.953	0.978	0.977

As shown in Table 13, while the NLI baseline and Qwen models performed well, GPT-5 achieved a near-perfect alignment with human annotators ($\kappa = 0.992$). Therefore, we selected GPT-5 as the primary LLM judge for evaluating consistency.

G.4 Prompt Template for Consistency

```
Given a reasoning text, infer whether it supports the conclusion that
    the image is persuasive.

### Instructions
- Use only the provided reasoning text.
- Do not rely on external knowledge or unstated assumptions.

### Output
Answer with exactly one word:
- yes
- no
```

G.5 Prompt Template for Groundedness

```
You are a QA auditor evaluating the "Visual Groundedness" of a
    reasoning text against an image.

You will be given:
1. An image
2. A reasoning text about the image

Your objective is to perform a Groundedness Evaluation. You must
    determine if the concrete visual claims made in the reasoning text
     physically exist in the provided image.

### Evaluation Rules:
1. Evaluate Concrete Visual Claims ONLY: Mentally break down the
    reasoning text into physical visual claims (objects, characters,
    gestures, backgrounds, explicit text). DO NOT evaluate subjective
    evaluations, viewer impacts, or rhetorical conclusions (e.g.,
    ignore statements like "The design lacks clarity", "This is
    unpersuasive", or "It compels the viewer").
2. "Message" vs. "Physical Text": If the reasoning says "The image
    conveys the message 'Do not smoke'", do NOT look for a physical
    sign saying "Do not smoke". Treat it as a thematic statement. ONLY
     strictly verify text if the reasoning explicitly claims it is
    written on a physical object (e.g., "The sign reads 'X'").
3. Lenient Text Matching for AI Images: AI-generated images often
    contain mangled text. If the reasoning quotes text and the image
    contains text that is clearly attempting to spell that phrase (e.g
    ., "EMERGENCY XIT", "Mid-d'y"), you MUST consider the claim
    GROUNDED. Do not penalize minor typos, missing letters, or strange
     characters.
4. Contextual Synonyms & Reasonable Assumptions: Use reasonable human
    logic. Accept "student" for someone studying. If a claim mentions
    "empty bottles" and the bottles are opaque, assume they are empty
    based on context. Be lenient with quantity descriptors like "
    filled," "abundance," or "scattered."
5. STRICT Relational and State Accuracy: You must verify the specific
    *state*, *adjectives*, and *relationships* of objects, not just
    their presence. If the text claims a "stark contrast between a
    dirty hood and the rest of the kitchen," and the rest of the
```

```
    kitchen is ALSO dirty, the claim is false. If it claims chains are
     "broken" but they are intact, or "scattered money" but they are
    just blank papers, the claim is false.
6. Zero Tolerance for Physical Hallucinations: If the reasoning
    describes EVEN ONE concrete physical element, object, specific
    state, or character that is completely NOT visible in the image,
    the final label MUST be "No". If all concrete physical claims and
    their described states are perfectly visible, the label MUST be "
    Yes".

### Output Format:
Return a JSON object ONLY. Do not include markdown formatting or
    explanations outside the JSON.

{{
"label": "Yes" or "No"
}}

### Input:
- Reasoning text:
{reason_text}
```

G.6 Prompt Template for Image Editing

```
You are given a message, an image based on that message, and an
    explanation describing why the image is persuasive or unpersuasive
    .

Your task is to identify the single visual element in the explanation
    that should be edited so that the image conveys the opposite
    judgment.

### Definition
Visual element is the single visual factor in the image that is most
    causally responsible for the conclusion described in the
    explanation.
It is the fundamental visual element in the image that drives the
    conclusion, rather than a secondary detail, descriptive attribute,
     or a strengthened version of an existing element.

### Steps
1. Read the image and the explanation carefully.
2. Identify the visual element that is most decisive for the
    explanation's conclusion.
3. Generate a concise image editing prompt that changes this element
    so that the image moves against the explanation.

### Rules
1. Remove
- If the visual element is a concrete object, person, or other visible
     entity that already appears in the image, remove it.
2. Modify
- If the visual element is atmosphere, setting, action, or emotion,
    modify it in the opposite visual direction.
- Select Modify only if removing the element would make the image
    unnatural.
3. Add
- If the visual element is the absence of something, add that missing
    element naturally.
```

```
### Rules
- Focus on one clear, editable visual target.
- The edit should be specific, visible, and directly reversible with
    respect to the explanation.
- The final prompt should be simple and precise.

### Remove Example
- Visual element: police cars and crime scene tape
- Good example: Remove the police cars and crime scene tape.

### Modify Example
- Visual element: excited people
- Good example: Modify the facial expressions of the people so that
    they look sad.
- Bad example: Modify the facial expressions of the people so that
    they look neutral.

### Modify Example
- Visual element: vibrant colors
- Good example: Modify the colors of the table and wall to make them
    dark and muted.
- Bad example: Modify the vibrant colors to be dull and less vivid.

### Contrast Example
- Visual element: the visual contrast between the solitary figure and
    the lively group
- Good example: Remove the lively group of people and fill the empty
    space naturally.
- Bad example: Reduce the visual contrast between the solitary figure
    and the lively group scene to make them appear more similar.

### Add Example
- Visual element: lack of books
- Good example: Add books on the desk.
- Bad example: Remove the books.

### Output
Return JSON:
{{
    "editable_target": "...",
    "edit_operation": "Add | Remove | Modify",
    "desired_change": "...",
    "final_prompt": "One simple sentence"
}}

### Input
- Message: {}
- Image: the input image
- Explanation: {}
```

G.7 Human Validation for Image Editing Pipeline

Figure 5 shows the annotation interface used to validate whether the image edits shift persuasiveness in the intended direction, a necessary condition for the sensitivity metric. We randomly sample 30 non-overlapping instances from each of three setups (Base-Reasoning, GRPO, and Reasoning-SFT) for student models, yielding 180 original–edited image pairs. We additionally include 20 control pairs with no meaningful difference in persuasiveness. Three annotators view each pair side by side and select which image is more persuasive for the given message (left, right, or ambiguous). An edit is considered successful when the judgment shifts counter to the model's predicted label: the edited image is preferred when the model predicts 'unpersuasive', whereas the original image is preferred when it predicts 'persuasive'. All 20 control pairs are correctly identified. Among the 180

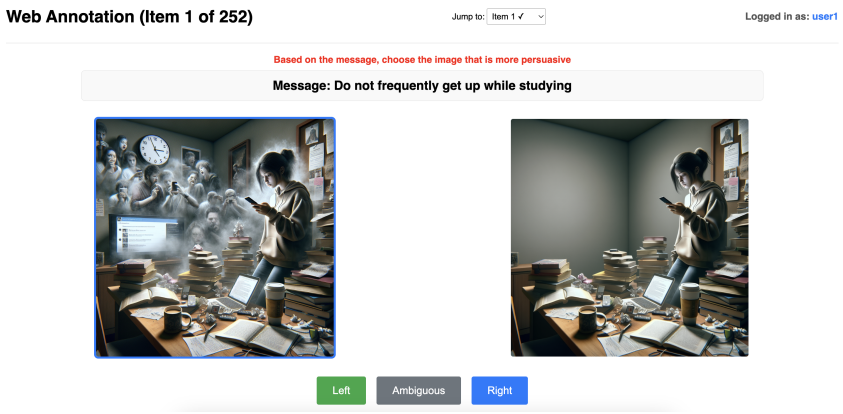


Figure 5: The annotation interface used for validating image editing pipeline.

edited pairs, five result in 1–1–1 ties; of the remaining 175 pairs, 131 (74.9%) are majority-judged to shift persuasiveness in the intended direction, with substantial inter-annotator agreement (Fleiss' $\kappa = 0.693$).

The three annotators were volunteer participants who provided informed consent and were compensated at the minimum wage of the country where the annotation was conducted. The task involved viewing persuasive advertisement images and making comparative judgments, posing no more than minimal risk to participants. No personally identifiable information was collected or retained, and participants were informed of the persuasive nature of the stimulus images prior to participation. This study is exempt from IRB approval under 45 CFR 46.104(d)(2)(i).

H Image-Editing Control Experiments

Sensitivity is intended to measure whether modifying the visual evidence referenced in a rationale affects the model's prediction. To verify that the measured sensitivity is not simply caused by unsuccessful edits or by any visible change to the image, we conduct two additional control experiments. First, we apply a random edit to each image without providing the rationale or target message to the editing model. Second, we apply an irrelevant but visible edit that is designed to preserve the original persuasiveness of the image.

Table 14: Sensitivity under different image-editing controls.

Family	Setup	Sensitivity	Random Edit	Irrelevant Edit
Qwen	Base-Reasoning	0.821	0.522	0.531
	GRPO-Ans	0.690	0.502	0.502
	Reasoning-SFT	0.700	0.556	0.526
Phi	Base-Reasoning	0.754	0.507	0.512
	GRPO-Ans	0.538	0.163	0.464
	Reasoning-SFT	0.664	0.485	0.517
Gemma	Base-Reasoning	0.635	0.572	0.543
	GRPO-Ans	0.564	0.367	0.439
	Reasoning-SFT	0.693	0.482	0.414
LLaVA	Base-Reasoning	0.795	0.536	0.589
	GRPO-Ans	0.577	0.476	0.443
	Reasoning-SFT	0.571	0.368	0.412

Across all evaluated setups, rationale-conditioned edits yield higher sensitivity than both random and irrelevant edits. On average, sensitivity is 0.667 for rationale-conditioned edits, compared with

0.461 and 0.491 for the two controls, respectively. This supports that the metric reflects changes to rationale-relevant visual evidence rather than a general response to image modification.

I Rationale Faithfulness Results

I.1 Quantitative Results

Table 15 reports the exact faithfulness scores corresponding to Figure 5. For R-I Groundedness, Reasoning-SFT achieves the highest model-averaged score of 78.7%, compared with 68.4% for GRPO-Ans and 64.3% for Base-Reasoning. The gains are particularly pronounced for Gemma and LLaVA, where Reasoning-SFT improves over Base-Reasoning by 26.2 and 21.4 percentage points, respectively. For R-D Consistency, Reasoning-SFT performs best across all four student models, reaching an average of 99.1%.

R-D Sensitivity shows a different pattern. Base-Reasoning achieves the highest sensitivity across all four models, with an average of 75.1%, while Reasoning-SFT obtains 65.7% and GRPO-Ans 59.2%. Thus, rationale supervision substantially improves groundedness and consistency, but does not uniformly improve sensitivity to interventions on rationale-relevant visual evidence, indicating a trade-off across the faithfulness dimensions.

Table 15: Rationale faithfulness across student models. Results are reported as mean $\pm$ standard deviation over five random seeds.

Metric	Setup	Qwen2.5-VL-7B	Phi-3.5-Vision	Gemma-3-4B	LLaVA-1.6-7B	Average
R-I Groundedness	Base-Reasoning	76.1 ± 4.6	70.6 ± 1.6	54.0 ± 2.5	56.4 ± 2.3	64.3 ± 1.7
	GRPO-Ans	83.2 ± 3.2	75.3 ± 3.3	56.9 ± 3.1	58.1 ± 2.5	68.4 ± 0.9
	Reasoning-SFT	85.2 ± 3.6	71.7 ± 2.4	80.2 ± 3.6	77.8 ± 9.8	78.7 ± 4.2
R-D Consistency	Base-Reasoning	96.0 ± 0.7	97.2 ± 1.6	83.7 ± 1.0	84.2 ± 5.4	90.3 ± 1.7
	GRPO-Ans	96.2 ± 3.6	96.0 ± 3.4	89.6 ± 1.7	85.3 ± 3.1	91.8 ± 1.3
	Reasoning-SFT	99.1 ± 0.6	98.2 ± 0.7	99.4 ± 0.6	99.7 ± 0.4	99.1 ± 0.2
R-D Sensitivity	Base-Reasoning	82.1 ± 4.5	75.4 ± 1.9	63.4 ± 1.6	79.5 ± 2.2	75.1 ± 1.6
	GRPO-Ans	68.9 ± 6.4	53.8 ± 1.7	56.4 ± 2.0	57.7 ± 2.0	59.2 ± 1.1
	Reasoning-SFT	70.0 ± 4.4	66.5 ± 2.0	69.3 ± 1.6	57.2 ± 2.2	65.7 ± 1.3

I.2 Qualitative Results

I.2.1 Rationale-to-image groundedness.

Figure 6(a) compares rationales generated for an image paired with the message "Prepare for national emergencies" (gold label: *No*). Base-Reasoning and Reasoning-SFT both identify the central object as a box marked with a red cross (i.e., a first-aid kit), whereas GRPO-Ans misidentifies it as a "gift box." Despite this ungrounded detail, GRPO-Ans still predicts the correct label because its conclusion relies on the broader absence of cues indicating national emergency preparedness. This example illustrates how answer-only reward can reinforce rationales containing ungrounded visual details as long as they lead to the correct prediction. More broadly, 27 of the 31 GRPO-Ans rationales judged ungrounded in this setting still yield correct predictions. Figure 6(b) shows a similar pattern: GRPO-Ans claims that the image contains no visual elements related to social media interactions, overlooking the man's engagement with his phone, which Reasoning-SFT explicitly identifies as a message-relevant visual cue.

I.2.2 Rationale-to-decision consistency.

Figure 7(a) compares rationales for the message "Do not place flammable items near the gas stove" (gold label: *No*). Although all three setups describe the same scene, they differ in whether the rationale supports the final prediction. Base-Reasoning argues that the image is persuasive and predicts *Yes*, yielding a consistent but incorrect prediction. GRPO-Ans similarly argues that the image is persuasive, but predicts *No*. Its prediction is therefore correct but inconsistent with its own rationale. Because GRPO-Ans is rewarded only for the final answer, such rationale–prediction



Message: Prepare for national emergencies **Gold:** *No*

Setup	Rationale (excerpt)	Grd.
Base-Reasoning	. . . a diverse group of people gathered around a person in a uniform holding a box with a red cross . . .	✓
GRPO-Ans	. . . the focus on a police officer holding a gift box and a diverse group of people standing around does not immediately suggest any connection to national emergencies.	×
Reasoning-SFT	. . . a group of people gathered around a police officer holding a first aid kit, which does not fully represent the broader concept of national emergency readiness.	✓

(a)



Message: Do not fight with others on social media **Gold:** *No*

Setup	Rationale (excerpt)	Grd.
GRPO-Ans	The image depicts a serene and peaceful scene with a person enjoying nature . . . The lack of any visual elements or context related to social media interactions makes it ineffective . . .	×
Reasoning-SFT	The image depicts a man sitting calmly in a serene park setting, seemingly engaged with his phone, which does not directly relate to the act of fighting on social media.	✓

(b)

Figure 6: Qualitative examples of rationale-to-image groundedness. Ungrounded content is highlighted in red, and Grd. denotes the groundedness judgment. In both examples, GRPO-Ans produces an ungrounded rationale while still arriving at the correct prediction.

inconsistency is not directly penalized: 17 of the 21 inconsistent GRPO-Ans samples in this setting still yield correct predictions. In contrast, Reasoning-SFT argues that the image is not persuasive and predicts *No*, producing a rationale consistent with the correct prediction. Figure 7(b) shows the same pattern: GRPO-Ans provides several reasons for why the image is persuasive before predicting *No*, whereas Reasoning-SFT provides a rationale that supports its *No* prediction.

I.2.3 Rationale-to-decision sensitivity.

Figure 8 presents an example for the message "Use an exercise ball to balance your body" (gold label: *Yes*). The original image shows two figures in the same seated posture: a grayscale figure on a chair and a colorful, anatomically detailed figure on an exercise ball. All three setups refer to the contrast between the two figures, but differ in the visual evidence they emphasize. Base-Reasoning and GRPO-Ans focus primarily on the contrast between the chair and the exercise ball. However, neither setup is sensitive to the removal of its cited evidence. In contrast, Reasoning-SFT identifies the colorful, anatomically detailed figure on the ball as key evidence supporting the message. When this figure is replaced with an ordinary person while the exercise ball is preserved, its prediction changes from *Yes* to *No*, indicating sensitivity to the cited evidence.



Message: Do not place flammable items near the gas stove **Gold:** *No*

Setup	Rationale (excerpt)	Cons.
Base-Reasoning	. . . The image effectively conveys the message by showing a person engaging in a potentially hazardous activity . . . making the message more persuasive.	Yes
GRPO-Ans	. . . The image effectively conveys the message by showing a potential hazard . . . This makes it persuasive in communicating the importance of keeping flammable items away from the gas stove.	No
Reasoning-SFT	The image fails to be persuasive because it shows a person safely interacting with a candle near a gas stove without any visible danger . . .	Yes

(a)



Message: Get a health check-up regularly **Gold:** *No*

Setup	Rationale (excerpt)	Cons.
GRPO-Ans	. . . The message conveyed by the image is indeed persuasive for several reasons: . . . Overall, the image effectively communicates the message of the importance of regular health check-ups . . .	No
Reasoning-SFT	The image fails to effectively convey the message . . . because it depicts a doctor looking overwhelmed and stressed, which may discourage rather than encourage viewers . . .	Yes

(b)

Figure 7: Qualitative examples of rationale-to-decision consistency. Cons. indicates whether the rationale supports the model's final prediction. Inconsistent cases are highlighted in red. In both examples, GRPO-Ans predicts the correct label despite its rationale supporting the opposite decision.



Original

Message: Use an exercise ball to balance your body **Gold:** *Yes*

Setup	Edited	Rationale (excerpt)	Pred.	Sens.
Base-Reasoning		. . . The contrast between the two figures further highlights the importance of the exercise ball in achieving **a more balanced posture**.	Yes→Yes	No
GRPO-Ans		. . . the person on the right sitting on **the exercise ball** . . . The contrast with the person on the left, who is sitting on a chair and appears to be in a more static position . . .	No→No	No
Reasoning-SFT		*Positive:* **The colorful, anatomically detailed figure on the ball**: The vibrant colors and focus on anatomy highlight the benefits of using an exercise ball . . .	Yes→No	Yes

Figure 8: Qualitative example of rationale-to-decision sensitivity. The original image is shown on the left, while Edited shows the counterfactual image obtained by removing or altering the visual evidence cited in each setup's rationale. Bold text marks the rationale evidence targeted by the corresponding edit. Pred. reports the prediction before and after the intervention, and Sens. indicates whether removing the cited evidence decreases the logit of the original prediction.