FROM PROMPT ALIGNMENT TO INTENT RECOVERY: RETHINKING TEXT-TO-IMAGE EVALUATION

Anonymous authors
Paper under double-blind review

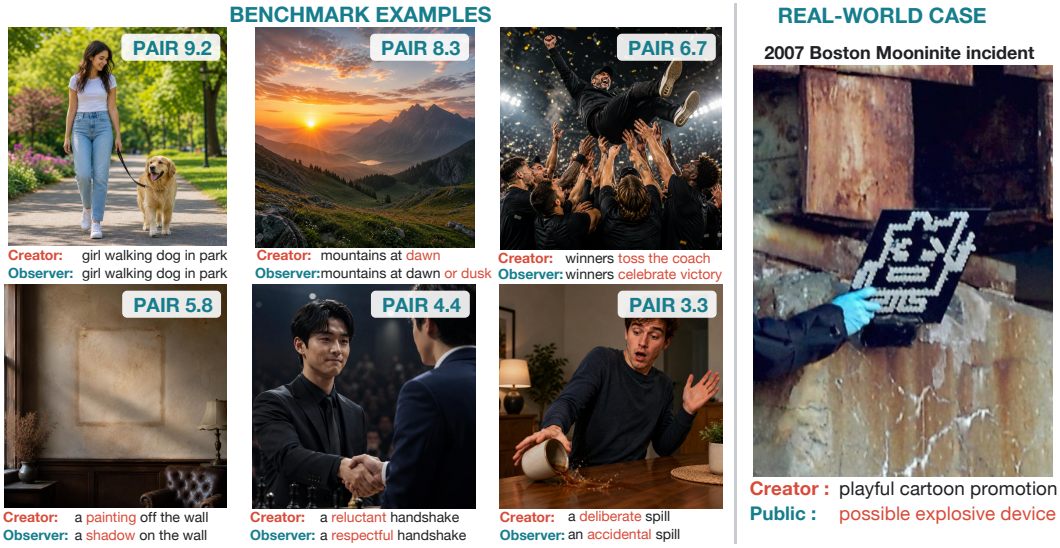

Figure 1: **benchmark and Real-world examples of Prompt-Agnostic Intent Recovery (PAIR).** Left: benchmark examples showing the creator's intent and the observer's interpretation of the image alone. Higher PAIR scores indicate closer agreement between the two. Right: a real-world analogue from the 2007 Boston Mooninite incident, where a promotional display was interpreted as a possible explosive device. Red highlights mark key mismatches between intended and recovered meaning.

ABSTRACT

Text-to-image models have made substantial progress in faithfully rendering creators' prompts. However, ambiguous visual cues may lead observers to misinterpret the creator's intent from images alone, potentially triggering public alarm, reputational damage, and economic losses. Existing prompt-conditioned metrics primarily evaluate creator-side prompt alignment but do not directly measure observer-side intent recovery. Hence, we introduce *Prompt-Agnostic Intent Recovery* (PAIR), a protocol for evaluating intent recovery from generated images alone. A prompt-agnostic observer produces an open-ended image description and a separate scorer evaluates the description against a pre-specified intent rubric. We construct a benchmark of 300 intents and 1,800 images from two generators. Across the benchmark, prompt alignment and intent recovery induce different candidate rankings. When selecting among six images per intent from two generators, PAIR-based selection achieves 9–17% higher recovery on held-out evaluations than CLIP, VQAScore, and a target-aware VLM selector. In diagnostic selector-disagreement cases, participants without prompt access achieved 85.9% two-choice accuracy for PAIR-selected images versus 39.4% for alignment-selected images. Finally, PAIR-guided refinement improves the mean

Code and data: https://anonymous.4open.science/r/viscomm-release-A078

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.

final PAIR score by 23% over repeated uncalibrated generation. These results identify observer-side intent recovery as an evaluation target not captured by prompt-conditioned alignment.

1 INTRODUCTION

Text-to-image generation is increasingly used in applications such as advertising, social media, and public information, where generated images can have substantial social and economic impact. Although recent text-to-image models have made substantial progress in faithfully rendering creators' prompts, observers typically do not have access to the source prompt and must infer meaning from visual evidence alone. When that evidence supports multiple plausible interpretations, the meaning recovered by an observer may differ from what the creator intended to convey. Such mismatches can weaken communication and, in consequential settings, contribute to public alarm, reputational harm, or economic loss. Figure 1 illustrates this problem with both a real-world case and generated examples. In the 2007 Boston Mooninite incident, an LED display intended to promote the animated series *Aqua Teen Hunger Force* was interpreted as a possible explosive device, triggering a major security response and transportation disruption (Slane, 2017). Text-to-image generation faces the same underlying gap: for example, an image intended to depict a deliberate spill may clearly show a fallen cup and spilled liquid, yet still be interpreted by an observer as an accident.

Existing text-to-image metrics do not directly evaluate this observer-side gap. Prompt-alignment, compositional-fidelity, and preference metrics typically assess generated images with the target text available (Radford et al., 2021; Hu et al., 2023; Lin et al., 2024; Xu et al., 2023; Kirstain et al., 2023). They therefore measure how well an image satisfies a known creator-side target, rather than what meaning an observer recovers from the image alone. We formulate this missing evaluation target as *Prompt-Agnostic Intent Recovery* (PAIR): whether an observer can recover a creator's pre-specified intent from a generated image without access to the source prompt.

To measure PAIR, a prompt-agnostic observer produces an open-ended description from the generated image alone, without access to the source prompt, creator intent, or scoring rubric. A separate target-aware scorer then evaluates the description against a pre-specified intent-specific rubric. This information boundary measures spontaneous intent recovery rather than target-conditioned recognition.

We construct a benchmark of 300 intents and 1,800 images from two generators, with pre-specified rubrics covering diverse forms of intended meaning. Across the benchmark, prompt alignment and intent recovery induce different candidate rankings. In six-image cross-generator pools, PAIR-based selection achieves approximately 9–17% higher held-out recovery than CLIP, VQAScore, and a target-aware VLM selector.

These differences persist across the tested observer families and under open-weight observer and scorer alternatives. Human studies provide external grounding: in diagnostic selector-disagreement cases, participants without prompt access identify the intended meaning in 85.9% of trials with PAIR-selected images versus 39.4% with alignment-selected images. Finally, PAIR-guided refinement increases the mean final PAIR score from 7.12 to 8.77, a 23% relative improvement over repeated uncalibrated generation. Together, these results establish observer-side intent recovery as a distinct evaluation target for selecting and refining generated images.

We make three contributions:

1. **A critical but overlooked evaluation problem.** We formulate Prompt-Agnostic Intent Recovery (PAIR): whether a generated image allows an observer to recover a creator's pre-specified intent without access to the source prompt. This captures an evaluation target not directly assessed by prompt-conditioned alignment.
2. **A benchmark and measurement protocol.** We construct a 300-intent benchmark spanning seven intent families, with 1,800 generated images and pre-specified signal rubrics. We measure intent recovery by separating prompt-agnostic open-ended observation from target-aware scoring.
3. **A systematic empirical evaluation.** Prompt-conditioned alignment scores and PAIR scores rank images differently, and PAIR-based selection gains persist across the tested generators,

observers, and scorers. Human studies validate the scorer and reveal differences in intent identification on diagnostic cases of selector disagreement. In the refinement study, generation using observer feedback achieves higher recovery scores than repeated uncalibrated generation.

2 Related Work

Prompt-conditioned text-to-image evaluation. Text-to-image generation spans autoregressive and diffusion families (Ramesh et al., 2022; Saharia et al., 2022; Yu et al., 2022; Podell et al., 2024; Esser et al., 2024; Betker et al., 2023). Most evaluation, however, measures an output relative to a known target text. Embedding-based metrics such as CLIP-based similarity measure image–text similarity (Radford et al., 2021), while compositional benchmarks and VQA-based methods test whether specified objects, attributes, and relations appear in the image (Cho et al., 2023; Ghosh et al., 2023; Hu et al., 2023; Lin et al., 2024; Li et al., 2024; Grimal et al., 2024). Broader suites and human-annotated benchmarks study the coverage and reliability of prompt-conditioned evaluation (Yarom et al., 2023; Wiles et al., 2025; Han et al., 2026; Zhang & Wan, 2025). These methods address an important target-conditioned question: given the prompt, how well does the image satisfy it? PAIR changes the information available to the observer at evaluation time. The observer receives neither the source prompt, the creator intent, nor the scoring rubric; only after an open-ended observation does a separate target-aware scorer evaluate intent recovery.

Preference, visual communication, and observer variation. Human evaluation of generated images also measures preference, quality, and communicative effectiveness. Pick-a-Pic and ImageReward collect or learn pairwise preferences over generated outputs (Kirstain et al., 2023; Xu et al., 2023), while human-evaluation work studies richer feedback formats and image-quality judgments (Petsiuk et al., 2022; Otani et al., 2023; Liang et al., 2024). Broader evaluation frameworks measure multiple dimensions of generated-image quality and semantic consistency (Lee et al., 2023b; Ku et al., 2024; Kamath et al., 2025; Li et al., 2026), and human-feedback alignment optimizes generators against preference signals (Lee et al., 2023a). Recent work further considers cultural and observer-dependent variation in image judgments (Malakouti et al., 2026; Rastogi et al., 2025). AIGI-VC is particularly relevant because it treats visual communication itself as an evaluation target, assessing information clarity and emotional interaction in advertising images (Tian et al., 2025). We operationalize communicative effectiveness as recovery of a pre-specified creator intent from the image alone, with that intent withheld from the observer. Preference, perceived quality, and observer variation may affect recovery but are not themselves the endpoint.

Pragmatic communication, intent understanding, and refinement. The broader distinction between intended and received meaning is longstanding in art and communication theory (Barthes, 1977; Eco, 1989). PAIR is conceptually related to pragmatic speaker–listener models, which define communication relative to what a recipient can infer from an utterance or signal (Andreas & Klein, 2016; Clark, 1996). Recent VLM work studies the recovery of intentions, perspectives, and mental states from visual inputs (Gao et al., 2025; Al Nazi et al., 2025; Lee et al., 2025). Unlike language-based intent recognition, PAIR evaluates generated images rather than observer capability: the generator mediates between the creator's verbalized intent and the visual evidence available to the observer. The VLM-as-judge methods evaluate outputs against known criteria (Lee et al., 2024; Li et al., 2025), while PAIR separates a prompt-agnostic observer from a target-aware scorer. Related generation systems address intent uncertainty before or during generation through clarification or user–model alignment (Hahn et al., 2025; Hegemann et al., 2026); PAIR assumes that the verbalized intent is already fixed and evaluates what remains recoverable after generation. Finally, iterative feedback and prompt optimization have been used to refine generated outputs (Madaan et al., 2023; Zhang et al., 2025; Ye et al., 2025). We further study PAIR as a feedback signal in a downstream image-refinement pipeline.

3 A Protocol for Prompt-Agnostic Intent Recovery

Let m be a creator's verbalized intent, $x \sim G(\cdot \mid m)$ a generated image, and O an observer that receives x but not m, the generation prompt, or the scoring rubric. The observer writes an open-

ended description $d = O(x)$. A separate scorer S receives d and a pre-specified, intent-specific signal rubric R_m, constructed before image evaluation (Figure 2).

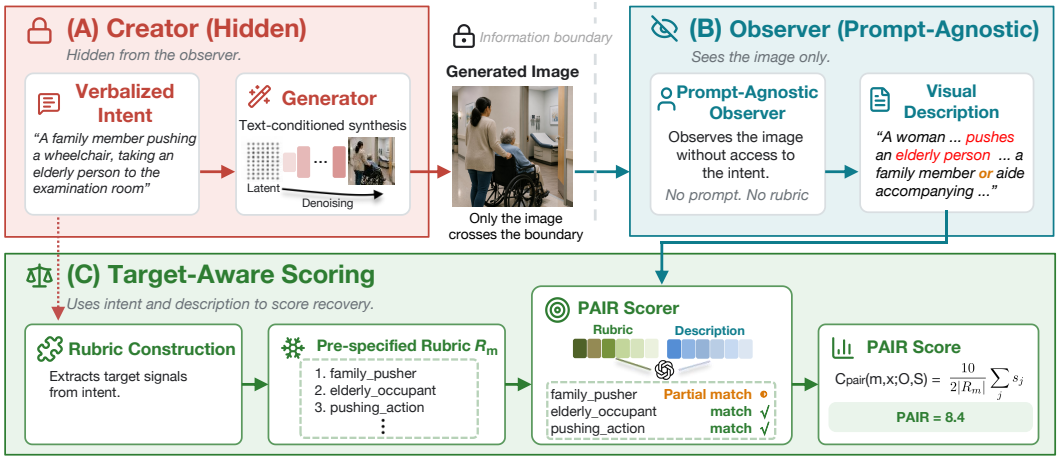

Figure 2: **PAIR separates prompt-agnostic observation from target-aware scoring.** (A) The creator verbalizes an intent and generates an image; only the generated image is passed to the observer. (B) The prompt-agnostic observer returns an open-ended description without access to the source prompt, creator intent, or rubric. (C) A separate target-aware scorer evaluates the description against a pre-specified, intent-specific rubric R_m.

The PAIR channel and its score are

$$m \longrightarrow G \longrightarrow x \longrightarrow O(x) = d, \qquad C_{\text{PAIR}}(m, x; O, S) = \frac{10}{2|R_m|} \sum_{j \in R_m} s_j, \tag{1}$$

where $s_j = S(d, r_j) \in \{0, 0.5, 1, 1.5, 2\}$ records how fully the description recovers signal j. Prompt withholding prevents target-conditioned visual search by the observer; the scorer can still evaluate recovered content offline.

Table 1: The graded-recovery contract on one pre-specified intentionality signal (*a person knocking over a cup on purpose*). Only the recorded observer reading changes across rows.

Observer reading	**Interpretation**	**Score**
"deliberately knocked the cup over"	recovers the intent	2
"may have knocked it over on purpose, although it could be accidental"	hedged between readings	1
"accidentally knocked the cup over"	contradicts the intent	0

The grading is deliberately recovery-graded rather than mention-based. Full credit on a signal requires the description to state the intended meaning clearly and correctly; hedged, either-or, or misattributed readings earn only partial credit, and naming the target as one of several possibilities counts as identifying the scene without recovering the intent. Table 1 illustrates the contract on one pre-specified intentionality signal. Throughout, PAIR evaluates recovery of the intended communicative signals, rather than general factual precision or stylistic quality of the whole description. Signals are equally weighted without outcome-tuned weights, and each rubric includes the central intent signal.

The prompt-agnostic observer receives one image and a fixed user prompt. Here, *prompt-agnostic* means that the observer does not receive the source generation prompt or creator intent; it still receives a fixed, target-independent observation instruction:

```
Look carefully at this image and describe it:  what do you
see, what is happening, and what - if anything - it seems to
convey (relationships, intentions, mood, story).  Report
```

```
everything you can determine from the image alone, in no
more than 5 sentences.
```

A matched Chinese translation is used in the human open-ended study, with the same five-sentence cap across observers. The score is indexed by observer O, scorer S, and rubric R_m; main analyses use GPT-5.5 as O and DeepSeek-V4-Pro as S. "Creator" denotes the benchmark's verbalized intent, and recovery means that the observer description contains its pre-specified target signals. Full prompts, role access, rubric construction, and service configuration appear in Appendix A.

Information boundary. Prompt withholding is the condition that distinguishes PAIR from target-conditioned alignment. For a fixed observer and observation instruction, the observer receives X but not the creator intent M, so its description D depends on M only through the generated image: $M \to X \to D$. This creates an image-mediated communication channel $M \to X \to D$: the creator specifies an intent, the generator renders an image, and the observer reconstructs meaning from that image. PAIR measures the success of this channel. This information boundary defines a distinct evaluation target: Section 5 shows that prompt-conditioned alignment and PAIR can prefer different images from the same candidate pool. Generation prompts also prohibit rendered text to prevent textual leakage (Appendix A).

4 PAIR Across Intent Families

The benchmark contains 300 English verbalized intents organized into seven intent families: concrete control, visual ambiguity, narrative, counterfactual absence, abstract relation, layered emotion, and intentionality. These families capture qualitatively different demands on visual interpretation, including direct depiction, disambiguation among plausible readings, event understanding, inference from visible traces, relational interpretation, affective interpretation, and intention attribution. Their design draws on related work in scene perception, event cognition, social and relational perception, emotion perception, and intention attribution (Oliva & Torralba, 2006; Zacks et al., 2007; Malik & Isik, 2023; Barrett et al., 2011; Heider & Simmel, 1944), adapted here to the problem of recovering intended meaning from generated static images.

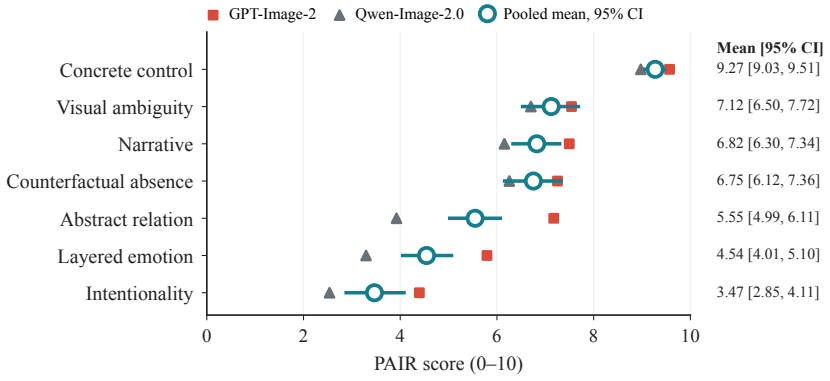

Figure 3: **PAIR varies across the seven intent families.** Open circles show pooled PAIR means with 95% bootstrap confidence intervals over intents; filled markers show generator-specific means for GPT-Image-2 and Qwen-Image-2.0. The two generators exhibit a similar broad intent-family ordering, although their absolute PAIR scores differ.

Thirty intents are concrete controls, which test direct depiction, while each of the remaining six families contains 45 intents. For each intent, we generate three images with GPT-Image-2 (OpenAI, 2026) and three with Qwen-Image-2.0 (Zhao et al., 2026), yielding 1,800 images in total. Each intent is paired with a pre-specified rubric containing two to five target signals used to evaluate recovery. Appendix B provides the full benchmark construction, intent-family definitions, and rubric design.

Figure 3 shows substantial variation in PAIR across the seven intent families. Concrete controls have the highest PAIR (9.27; 95% CI [9.03, 9.51]), whereas intentionality has the lowest (3.47

[2.85, 4.11]). Visual ambiguity, narrative, and counterfactual absence form a higher middle range (6.75–7.12), followed by abstract relation (5.55) and layered emotion (4.54). This intent-family profile shows that recoverability depends strongly on the kind of meaning an image must convey: directly visible content is recovered more reliably, whereas relational, affective, and intentional meanings require more informative visual evidence. The figure reports the full confidence intervals and generator-specific means. An ordinary-least-squares regression with intent-family indicators and controls for standardized prompt length, rubric size, and generator identity yields the same broad intent-family profile (adjusted means in Appendix C).

5 PAIR Reveals Selection Regret

For intent m, let $\mathcal{P}(m)$ denote a candidate pool and $T_q(m) \subseteq \mathcal{P}(m)$ the candidates tied for the highest score under selector q. Let x_q be sampled uniformly from $T_q(m)$, and define x_{PAIR} analogously for PAIR-based selection. Selection regret is the expected PAIR loss of selector q relative to PAIR-based selection:

$$\mathcal{R}(q) = \mathbb{E}_m[C_{\text{PAIR}}(m, x_{\text{PAIR}}; O, S) - C_{\text{PAIR}}(m, x_q; O, S)]\,, \tag{2}$$

Selection regret compares selectors on a common held-out PAIR endpoint using the same observer O and scorer S. CLIP, VQAScore, and a matched target-aware VLM selector only rank candidates; the latter uses the same description–rubric–scorer pipeline as PAIR but exposes the intent during description. Equation 2 therefore measures selector-induced loss under the PAIR construct, not general image quality; Section 7 separately tests the human consequences of these disagreements.

To avoid reusing noisy observer scores for both PAIR selection and evaluation, all regret estimates use split-sample evaluation: selections are made from two observer calls and evaluated on the held-out third, averaged over three folds, for PAIR-based and metric-based selection alike. A uniform random selector anchors the scale.

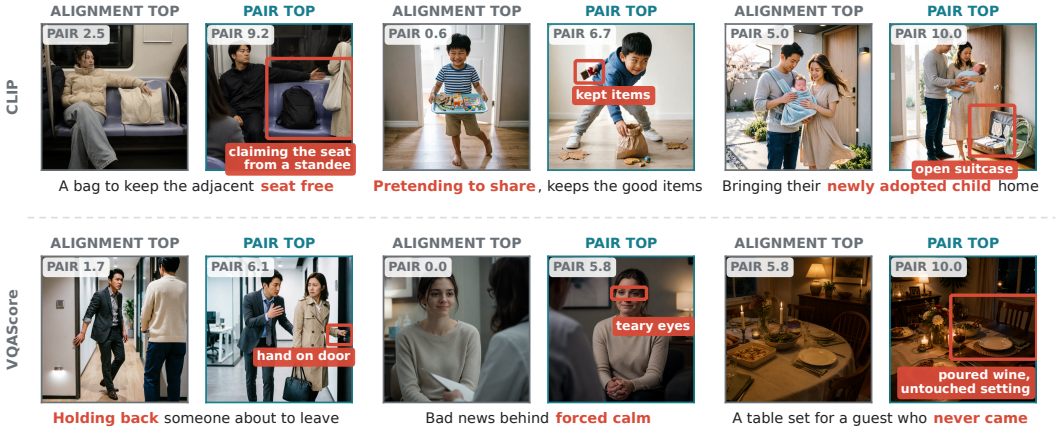

Figure 4: **Prompt alignment and PAIR can select different images from the same candidate pool.** For each intent, the alignment-selected image is shown on the left and the PAIR-selected image on the right. Corner labels give PAIR scores; red boxes highlight intent-relevant cues.

We first isolate candidate ranking within each generator using three-candidate pools, where a metric ranks candidates from one generator for one intent. PAIR-based selection yields a held-out mean of 6.67. CLIP, VQAScore, and the target-aware VLM score yield 6.19, 6.39, and 6.27, corresponding to regrets of 0.48 (95% CI [0.38, 0.59]), 0.28 [0.19, 0.37], and 0.40 [0.32, 0.48], respectively. A uniform random selector within each pool averages 6.07 (0.60 regret [0.52, 0.69]). Figure 4 gives concrete same-generator examples, and per-generator per-family cells remain positive in nearly every stratum (Table 6).

The six-candidate pooled analysis is a stronger decision setting that combines candidate ranking with cross-generator routing. PAIR-based selection yields a held-out mean of 7.50. CLIP selects

candidates averaging 6.41, for 1.09 regret [0.89, 1.30], with the loss reaching at least two points on 21.7% of intents. VQAScore selects 6.85, for 0.65 regret [0.46, 0.85], with losses of at least two points on 16.3%. The target-aware VLM score, which sees both image and intent, selects candidates averaging 6.52, for 0.97 regret [0.83, 1.13]. The random selector leaves 1.43 regret [1.27, 1.59] pooled. An image can therefore score highly under prompt-conditioned selection while providing weaker visual evidence for prompt-agnostic intent recovery.

6 PAIR Generalizes Across Evaluation Stacks

Do the main PAIR findings depend on the particular observer, scorer, or generator used in the primary analysis? We next evaluate PAIR across observer families, scoring models, and image generators. This analysis characterizes which properties of the endpoint transfer across evaluation stacks.

Held-out observer transfer. We first test whether PAIR-based selection transfers across observer families. Across the 300 intent pools, candidates selected using the GPT-5.5 prompt-agnostic observer are re-observed by Gemini-3.5-Flash, with the primary DeepSeek scorer and rubrics unchanged (Figure 6 in Appendix A.4). Relative to CLIP and VQAScore selection, PAIR-based selection gains 0.75 points (95% CI [0.52, 0.99]) and 0.42 [0.23, 0.63], respectively. Reversing the observer roles gives gains of 0.77 [0.56, 0.98] over CLIP and 0.33 [0.15, 0.52] over VQAScore. Selection gains transfer bidirectionally between GPT-5.5 and Gemini-3.5-Flash observers.

Open-weight observer. An open-weight observer, Qwen3-VL-32B-Instruct (Bai et al., 2025), re-observed all 1,800 images under the identical prompt-agnostic observer prompt, and scores were assigned using the same target-aware DeepSeek scorer and rubrics. Intent-level agreement with both closed observers is Spearman $\rho = 0.92$; the intent-family ordering replicates exactly. Selection gains also transfer in both directions: with the open observer evaluating GPT-5.5 selections, the gains over CLIP and VQAScore are 0.62 [0.35, 0.88] and 0.45 [0.19, 0.70]; reversing the roles gives 0.66 [0.45, 0.87] and 0.22 [0.02, 0.43]. Qwen3-VL reproduces the intent-family ordering and transfers PAIR-based selections in both directions. Protocol, intent-family profile, transfer detail, and score-distribution notes appear in Appendix C.5.

Open-weight scorer. An open-weight scorer, Qwen3-32B (Yang et al., 2025), re-scored all recorded GPT-5.5 observer descriptions (three per image) for the 1,800 benchmark images under the identical rubrics and normalization, with the observer and prompts held fixed. The open scorer agrees strongly with the primary DeepSeek scorer at both image level ($\rho = 0.896$) and intent level ($\rho = 0.935$). Its absolute scores are systematically higher (overall mean 6.65 versus 6.07, with the largest calibration difference on layered emotion), so scorer calibration is not identical; nevertheless, the broad intent-family profile remains similar, and selection regret over CLIP, VQAScore, and the target-aware VLM remains positive under the open scorer (0.86, 0.53, and 0.89). The primary observer and scorer come from different model families, and both roles replicate independently with open-weight alternatives. Details in Appendix C.6.

Human scorer validation. A separate five-rater study also validates the primary scorer against bilingual human judgments: across 352 description-signal items, scorer-human consensus agreement is $\kappa = 0.87$ [0.82, 0.91], and description-level human- and model-scored PAIR correlate at $\rho = 0.91$. Full protocol and sensitivity analyses appear in Appendix D.4.

Generator stability. Across the two generators, the seven intent-family means correlate at Spearman $\rho = 0.96$ and per-intent scores at $\rho = 0.65$: the broad intent-family profile is highly similar across generators, while agreement at the individual-intent level is more moderate.

Rubric specificity. Rubric swaps give 7.19 with the correct rubric, compared with 0.73 for same-family swaps and 0.37 for cross-group swaps. This large drop indicates that high scores require overlap with the intended signal set rather than merely a detailed description (details in Appendix C).

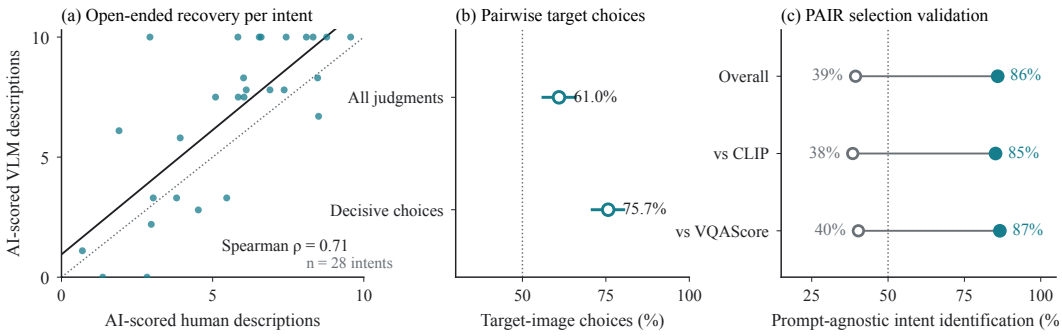

Figure 5: **Human studies provide complementary grounding for PAIR.** Left: human and VLM open-ended PAIR scores correlate across 28 intents under the same target-aware scorer. Middle: intent-informed preference for PAIR-guided refinement over uncalibrated generation. Right: prompt-agnostic intent identification on rule-selected selector-disagreement cases, comparing PAIR-selected with CLIP-/VQAScore-selected images.

7 Human Evaluation of Prompt-Agnostic Intent Recovery

We evaluate PAIR through five participant-facing human studies involving 249 participants and roughly 5,000 judgments and descriptions. Three main studies examine open-ended interpretation, prompt-agnostic intent identification under prompt-agnostic observation, and intent-informed image preference; two supplementary studies provide a four-way forced-choice identification test with 61 participants and an informed-rating test with 39 participants (Appendix D).

Open-ended human recovery. 49 participants wrote open-ended descriptions of generated images under matched prompt-withholding instructions (a Chinese translation of the observer instruction). The study covered 28 stratified intents split into two complementary halves; each intent contributes one benchmark image, and every human description was evaluated using the same prespecified rubrics and target-aware scorer used for the VLM descriptions. Across the resulting 28 intent-level means, human and VLM PAIR scores move together: Spearman $\rho = 0.715$ and Pearson $r = 0.751$ (Figure 5, left). The association remains after mean-centering both series within each intent family ($\rho = 0.720$). Human and VLM observers exhibit closely aligned intent-level recovery patterns, including within-family variation. Re-scoring with the open-weight scorer gives $\rho = 0.762$. Statistical tests and per-intent accounting appear in Appendix D.

Prompt-agnostic validation of selection regret. A second study asks whether the selection disagreements at the center of Figure 4 matter to uninformed people. We constructed a 28-pair diagnostic set using the quantitative selection rule described in Appendix D.3, with four pairs from each of the seven intent families and equal representation of CLIP and VQAScore comparisons. Each of the 53 participants viewed a single image per pair without the prompt and chose between two interpretations; participants never saw both selected images of the same pair, and image condition and option order were counterbalanced across two forms. The alternative option was derived from a recorded prompt-agnostic reading of the alignment-selected image. The test therefore compares the intended interpretation with the specific competing interpretation associated with that image. Intent-identification accuracy was 85.9% for PAIR-selected images and 39.4% for alignment-selected ones, a mean pair-level difference of $+46.6$ percentage points (two-way participant $\times$ pair bootstrap 95% CI $[+34.1, +59.0]$). PAIR-selected images led in 26 of 28 pairs, with similar gaps against CLIP and VQAScore selections (Figure 5, right). Because this is a rule-selected diagnostic set of selector disagreements, the estimate characterizes the consequence of disagreement rather than a benchmark-wide average. Pair- and participant-level tests and per-item results appear in Appendix D.

Intent-informed pairwise preference. A third main study tests the downstream PAIR-guided refinement pipeline with intent-informed raters: a 20-item pairwise evaluation comparing refined images against repeated uncalibrated generations from the original prompt. Across 940 judgments

from 47 participants, the refined image was preferred in 61.0% of all responses and in 75.7% of decisive A/B responses (Figure 5, middle). Full protocol and statistical tests appear in Appendix D.

8 PAIR AS A REFINEMENT SIGNAL

Beyond evaluation and candidate selection, we probe PAIR as a refinement signal over the same 300 intents using four GPT-Image-2 pipelines: structured PAIR feedback, self-guided refinement, text-only prompt rewriting, and repeated uncalibrated generation. In the structured-feedback condition, a description from the prompt-agnostic observer and the PAIR score guide a target-aware planner; the uncalibrated condition instead repeats generation from the original prompt without observer feedback.

Structured feedback improves final PAIR over uncalibrated generation by 1.65 points (8.77 vs. 7.12; 95% CI [1.41, 1.89]; paired $t = 13.5$), while self-guided refinement and text-only rewriting reach 7.98 and 6.66. The closer comparison to self-guided refinement holds the target-aware planner and iteration schedule fixed while removing the prompt-agnostic observation step, yielding a $+0.79$ advantage for structured feedback. Because these are complete pipeline configurations, the results characterize end-to-end effectiveness rather than isolate a single causal component; full details appear in Appendix E.

Re-evaluation with an independent open-weight observer–scorer stack (Qwen3-VL-32B and Qwen3-32B) retains a positive advantage of $+0.69$ points (7.97 vs. 7.28; 95% CI [0.45, 0.93]; paired $t = 5.7$). Together, these results support PAIR-derived observer feedback as an effective refinement signal across both evaluation stacks.

9 DISCUSSION

The intent-family profile suggests that public visual evidence is central to intent recovery: visible state changes, event traces, and conventional social configurations support recovery more readily than implicit motives or mixed affect. This pattern remains stable across generators, observers, and adjustments for prompt length and rubric size, showing where generated images communicate effectively and where stronger visual cues are needed. Together with the selection and refinement results, PAIR provides a practical observer-side objective for choosing and improving images intended to communicate without their source prompts.

10 LIMITATIONS

PAIR remains conditioned on the observer, scorer, rubric, and deliberate-inspection protocol; broader populations and scoring configurations may exhibit different calibration. Low scores may reflect generator limitations, intrinsic ambiguity in static images, or both, which the present data do not separate. PAIR also requires one observer and one scorer call per image, making it more costly than single-pass metrics; future work can extend PAIR across broader audiences and viewing contexts and evaluate matched refinement components.

11 CONCLUSION

We introduced Prompt-Agnostic Intent Recovery (PAIR) to evaluate how effectively a generated image communicates its creator's intended meaning without revealing the source prompt to the observer. Across 300 intents, two generators, and 1,800 images, PAIR reveals systematic variation in recoverability across intent families and identifies candidate-selection gaps left by prompt-conditioned metrics. These findings transfer across observer and scorer families, correspond with human intent recovery, and extend to PAIR-guided image refinement. By making observer-side recovery measurable, PAIR provides a unified objective for evaluating, selecting, and improving generated images for communication beyond the prompt.

AI USE STATEMENT

In this work, we used generative AI tools for refining hypotheses, reviewing experimental design, implementing and debugging methods, cleaning and reformatting data, interpreting results, and Chinese–English translation. We also used them for drafting and editing the manuscript, preparing figures and survey materials, and surveying the literature. We did not use AI to fabricate or alter any result: the experimental images are outputs of the evaluated text-to-image models, the human-study responses come from the reported participants, and the headline numerical results are recomputed from archived records. We verified AI-assisted work by author review of all retained text, testing AI-assisted code against the released analysis and verification scripts, and recomputing every reported number from the archived experiment records against a versioned evidence manifest. We have reviewed all AI-assisted work and take responsibility for the final content, including text, claims, code, and figures produced with the aid of generative AI.

ETHICS STATEMENT

The human-subjects components comprised online studies with lay participants. The landing page identified the study's academic purpose and voluntary participation, and participants gave consent by proceeding. Participants received a small payment of approximately 10 RMB. No demographic fields were collected; self-entered display names were used only for internal duplicate-participant screening and were excluded from the released artifacts. The released artifacts remove names, response identifiers, timing fields, and free-text responses, retaining only per-intent aggregates and de-identified choice matrices. Protocol and privacy details appear in Appendix D.

REPRODUCIBILITY STATEMENT

Appendix A specifies role-level information access, prompts, rubrics, model information, and aggregation rules. The supplementary package preserves the protocol artifacts: the intent rubrics, the observer prompts, the recorded observer descriptions, the PAIR scores from both observer families, the open-weight observer and open-weight scorer score files, trial timestamps, evidence and figure manifests, and scripts that rebuild the headline tables, figures, and de-identified human-study summaries. The package additionally preserves the scorer-validation item set, de-identified human signal ratings, and the analysis script that reproduces all agreement statistics. The submission verifier checks those artifacts against the manuscript, the compiled page boundaries, and the displayed image assets. Code, rubrics, the intent set, recorded descriptions, score files, and analysis scripts are released anonymously at https://anonymous.4open.science/r/viscomm-release-A078 for double-blind review.

REFERENCES

Zabir Al Nazi, GM Shahariar, Md. Abrar Hossain, and Wei Peng. Are vision language models cross-cultural theory of mind reasoners? *arXiv:2512.17394*, 2025.

Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. In *EMNLP*, 2016.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5), 2011. doi: 10.1177/0963721411422522.

Roland Barthes. The death of the author. *Image-Music-Text*, 1977.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *OpenAI Technical Report*, 2023.

Jaemin Cho, Abhay Zala, and Mohit Bansal. DALL-Eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*, 2023.

Herbert H. Clark. *Using Language*. Cambridge University Press, 1996.

Umberto Eco. *The Open Work*. Harvard University Press, 1989.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *ICML*, 2024.

Qingying Gao, Yijiang Li, Haiyun Lyu, Haoran Sun, Dezhi Luo, and Hokin Deng. Vision language models see what you want but not what you see. In *ICLR Workshop on Bidirectional Human-AI Alignment*, 2025.

Dhruba Ghosh et al. GenEval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023.

Paul Grimal, Hervé Le Borgne, Olivier Ferret, and Julien Tourille. TIAM—a metric for evaluating alignment in text-to-image generation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. doi: 10.1109/WACV57701.2024.00287.

Meera Hahn, Wenjun Zeng, Nithish Kannen, Rich Galt, Kartikeya Badola, Been Kim, and Zi Wang. Proactive agents for multi-turn text-to-image generation under uncertainty. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21591–21628. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/hahn25a.html.

Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and Chongyi Li. EvalMuse-40K: A fine-grained benchmark with comprehensive human annotations for text-to-image generation model alignment evaluation. In *AAAI*, 2026.

Lena Hegemann, Xinyi Wen, Michael A. Hedderich, Tarmo Nurmi, and Hariharan Subramonyam. ToMigo: Interpretable design concept graphs for aligning generative AI with creative intent. *arXiv:2602.05825*, 2026.

Fritz Heider and Marianne Simmel. An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2):243–259, 1944. doi: 10.2307/1416950.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, 2023.

Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. ImagenHub: Standardizing the evaluation of conditional image generation models. In *ICLR*, 2024.

Kimin Lee et al. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023a.

Phillip Y. Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *ICCV*, pp. 9241–9251, 2025.

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. Prometheus-Vision: Vision-language model as a judge for fine-grained evaluation. In *ACL Findings*, 2024.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, et al. Holistic evaluation of text-to-image models. In *NeurIPS*, 2023b.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. GenAI-Bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *EMNLP*, 2025.

Zejian Li, Qi Liu, Jiaman Pan, et al. SCoRE: Standardized human evaluation provides a reliable measure for semantic consistency of text-to-image generation. *International Journal of Computer Vision*, 2026. doi: 10.1007/s11263-026-02842-y.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *CVPR*, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *ECCV*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*, 2023.

Sina Malakouti, Boqing Gong, and Adriana Kovashka. Culture in action: Evaluating text-to-image models through social activities. In *ICLR*, 2026.

Manasi Malik and Leyla Isik. Relational visual representations underlie human social interaction recognition. *Nature Communications*, 14:7317, 2023. doi: 10.1038/s41467-023-43156-8.

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. doi: 10.1016/S0079-6123(06)55002-2.

OpenAI. GPT-Image-2 (model documentation). https://platform.openai.com/docs/models, 2026.

Mayu Otani et al. Toward verifiable and reproducible human evaluation for text-to-image generation. In *CVPR*, 2023.

Vitali Petsiuk et al. Human evaluation of text-to-image models on a multi-task benchmark. In *NeurIPS Workshop on Human Evaluation of Generative Models*, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *ICLR*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv:2204.06125*, 2022.

Charvi Rastogi, Tian Huey Teh, Pushkar Mishra, et al. Whose view of safety? a deep DIVE dataset for pluralistic alignment of text-to-image models. In *NeurIPS Datasets and Benchmarks*, 2025.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022.

Kevin Slane. Looking back at the boston mooninite panic, 10 years later. *Boston.com*, January 2017. URL https://www.boston.com/news/local-news/2017/01/31/looking-back-at-the-boston-mooninite-panic-10-years-later/.

Yu Tian, Yixuan Li, Baoliang Chen, Hanwei Zhu, Shiqi Wang, and Sam Kwong. AI-generated image quality assessment in visual communication. In *AAAI*, 2025.

Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Pinelopi Papalampidi, Ira Ktena, Christopher Knutsen, Cyrus Rashtchian, Anant Nawalgaria, Jordi Pont-Tuset, and Aida Nematzadeh. Revisiting text-to-image evaluation with Gecko: On metrics, prompts, and human ratings. In *ICLR*, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. In *NeurIPS*, 2023.

Wen Ye, Zhaocheng Liu, Yuwei Gui, Tingyu Yuan, Yunyue Su, Bowen Fang, Chaoyang Zhao, Qiang Liu, and Liang Wang. GenPilot: A multi-agent system for test-time prompt optimization in image generation. In *EMNLP Findings*, 2025.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *TMLR*, 2022.

Jeffrey M. Zacks, Nicole K. Speer, Khena M. Swallow, Todd S. Braver, and Jeremy R. Reynolds. Event perception: A mind-brain perspective. *Psychological Bulletin*, 133(2):273–293, 2007. doi: 10.1037/0033-2909.133.2.273.

Huixuan Zhang and Xiaojun Wan. Re-thinking the automatic evaluation of image-text alignment in text-to-image models. *arXiv preprint arXiv:2506.08480*, 2025.

Xinchen Zhang, Ling Yang, Guohao Li, Yaqi Cai, Jiake Xie, Yong Tang, Yujiu Yang, Mengdi Wang, and Bin Cui. IterComp: Iterative composition-aware feedback learning from model gallery for text-to-image generation. In *ICLR*, 2025.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, et al. Qwen-Image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

A Protocol and Reproducibility Details

A.1 Information Boundaries

Table 2 records request-time access. Each VLM perception request contains the fixed observer prompt and image, but no intent or rubric. After perception, the orchestrator uses intent metadata to retrieve the rubric and logs that metadata with the result; its `_history` is not injected into subsequent prompt-agnostic VLM requests. The scorer receives the recorded description rather than the image. Together, these controls implement the request-level information boundary used throughout the experiments.

Table 2: Role access in the PAIR protocol. A checkmark means that the role may access the field during that operation.

Role	Intent	Gen. prompt	Image	Description	Rubric
Creator/planner	✓	✓	✓	optional	✓
Image generator		✓			
prompt-agnostic observer			✓		
Target-aware scorer	✓			✓	✓

The *verbalized intent* is the benchmark message to be recovered. In the generation request it is followed by the fixed constraint

```
Do NOT include any text, letters, or numbers in the image.
```

This wrapper prevents the generator from communicating the target through rendered text rather than visual evidence.

A.2 Observer and Scorer Prompts

The prompt-agnostic observer receives one image and the following user prompt:

```
Look carefully at this image and describe it: what do you
see, what is happening, and what - if anything - it seems to
convey (relationships, intentions, mood, story). Report
everything you can determine from the image alone, in no
more than 5 sentences.
```

The instruction elicits an open-ended communicative reading, including relationships, intentions, mood, and narrative when supported by the image. A matched Chinese translation of this instruction is used in the human open-ended study. The target-aware observer used for the matched control selector receives the same image together with the verbalized intent:

```
You know the creator intended to convey: "[intent]".
Describe what you perceive in this image in relation to that
intent, in no more than 5 sentences.
```

Its description is scored against the identical rubric by the same scorer, so this selector differs from PAIR only in the information available at description time.

The scorer receives the pre-specified rubric and the observer text through the following template, with one block per target signal:

```
Score an observer description against the rubric below using
SEMANTIC matching, NOT literal word matching. Judge whether
the meaning required by each signal is present. Synonyms,
paraphrases, and equivalent expressions count. For each
signal, give 0, 0.5, 1, 1.5, or 2: 2 means clearly present
and correctly specified; 1 means partial or vague; 0 means
absent or wrong.
RUBRIC: [signal description; YES, PARTIAL, and NO criteria]
OBSERVER: [prompt-agnostic description]
Reply in JSON only: {"signals": {"<id>":
<0|0.5|1|1.5|2>}}.
```

The arithmetic normalization in Equation 1 is applied after the structured response. Missing signal IDs receive zero. Rubrics remain read-only during experiment execution, and missing rubrics are logged as integrity failures and excluded from the final datasets.

A.3 Model and Service Configuration

The benchmark uses square 1024×1024 images generated with GPT-Image-2 and Qwen-Image-2.0. The primary prompt-agnostic observer is GPT-5.5, the transfer observer is Gemini-3.5-Flash, and the text scorer is DeepSeek-V4-Pro. Benchmark trial timestamps span 2026-06-18 to 2026-06-19, and refinement trial timestamps span 2026-07-02 to 2026-07-16. Observation and scoring calls use temperature zero and an output cap of 4,096 tokens; the observer's five-sentence instruction is the effective length constraint. For the primary and transfer observers, three independent observer/scorer calls are averaged for each benchmark image; the open-weight observer replication retains the first successful call per image (Appendix C.5).

Repeated hosted-service responses were not byte-identical despite temperature zero. We therefore retain all valid repeated descriptions and average their scores. Model information, trial timestamps, descriptions, and scores are preserved to support aggregate reconstruction.

A.4 Observer Transfer Configurations

The primary configuration pairs the GPT-5.5 prompt-agnostic observer with DeepSeek-V4-Pro. The observer-transfer configuration replaces only the observer with Gemini-3.5-Flash; the intent rubrics, scoring model, prompts, and normalization remain fixed. Across intent means, the resulting PAIR scores have Spearman $\rho = 0.935$ (Pearson $r = 0.935$). At the finer trial level, within-intent Pearson agreement averages $r = 0.59$ over 276 intents for which the statistic is defined. These results show strong agreement at the aggregate level and meaningful observer variation at the individual-trial level. The transfer analysis covers all 300 intent pools and 600 same-generator groups. An open-weight replication that replaces the observer with Qwen3-VL-32B-Instruct under the same fixed scorer, rubrics, prompts, and normalization appears in Appendix C.5.

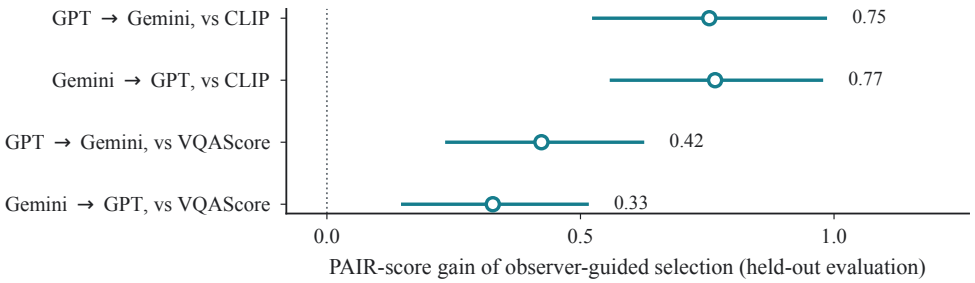

Figure 6: **PAIR-based selections transfer across prompt-agnostic observers.** Only the prompt-agnostic observer changes between GPT-5.5 and Gemini-3.5-Flash; the primary target-aware DeepSeek scorer and rubrics remain fixed. Points give the mean PAIR-score gain of observer-guided selection over CLIP or VQAScore selection under held-out evaluation, with 95% bootstrap intervals over 300 intent pools.

B Benchmark and Rubric Construction

B.1 Intent Sampling

The benchmark contains 300 English verbalized intents organized into seven intent families. The families were selected to cover qualitatively different forms of visual evidence involved in recovering meaning from a static image: direct depiction, disambiguation among plausible readings, event structure, inference from visible traces, relational interpretation, affective interpretation, and intention attribution. Their design draws on related work in scene perception, event cognition, social and relational perception, context-sensitive emotion perception, and intention attribution.

Table 3: Composition of the 300-intent PAIR benchmark.

Intent family	Primary visual-interpretation demand	Intents
Concrete control	Direct depiction of visible content	30
Visual ambiguity	Disambiguation among plausible scene readings	45
Narrative	Recovery of event or social-script structure	45
Counterfactual absence	Inference from visible traces of something absent or completed	45
Abstract relation	Recovery of relational or structural meaning	45
Layered emotion	Affective interpretation beyond surface expression	45
Intentionality	Attribution of the goal or motive underlying an action	45
Total		**300**

Concrete controls test recovery of directly visible objects, attributes, and actions. Visual-ambiguity items admit multiple plausible scene readings and require evidence that favors the intended one. Narrative items require recovery of an event or social script beyond isolated object presence. Counterfactual-absence items test whether visible traces make a missing, completed, or no-longer-present event recoverable. Abstract-relation items require recovery of relational or structural mean-

ing. Layered-emotion items require affective interpretation beyond the most immediate surface expression, while intentionality items require recovery of the goal or motive underlying an observable action.

Thirty intents are concrete controls, and each of the remaining six families contains 45 intents. Intent wording specifies the intended communicative meaning while avoiding a single prescribed visual realization. Table 3 summarizes the benchmark composition.

B.2 Signal Rubric Construction

Rubrics were model-generated before result scoring and reviewed by the authors. Every rubric was read by a human auditor; items with generic, trivially visible, or unfairly specific signals were revised or replaced before image scoring began. The authoring template requested two to five distinctive signals, excluded generic facts such as "a person is present," and required semantic YES/PARTIAL/NO criteria. Schema and integrity checks verify one rubric per intent, unique signal IDs, required fields, and the requested signal-count range. Rubrics were generated with DeepSeek-V4-Pro at temperature 0.0 (max tokens 4096) using a single authoring template. The template instructs the model to define 2–5 signals that constitute the creator's intended meaning, to exclude trivial facts generic to any image (e.g., "a person is present"), to make any hidden, inferred, or contrastive element the central signal, and to define YES(2)/PARTIAL(1)/NO(0) criteria around semantic equivalence with listed synonymous phrasings rather than exact wording. The full template is released with the benchmark code (`generate_rubrics.py`). The 300 rubrics contain 913 signals: 46 intents have two, 198 have three, 53 have four, and 3 have five (mean 3.04); 66.0% use exactly three. Content outside the rubric does not reduce the score. Rubric-swap and open-weight scorer analyses are reported in Appendices C.3 and C.6.

C Recoverability and Selection Analysis Details

C.1 Aggregation and Confidence Intervals

Each of the 300 intents has six generated images: three from each generator. The primary PAIR value for an image averages three observer/scorer returns. Intent-family means first average all available images within each intent and then average across intents, so each intent contributes equally to its family mean. Figure 3 uses percentile bootstrap intervals from 10,000 resamples of intents within each intent family, with fixed seed 20260729.

Table 4: Generator-stratified recoverability means recomputed from the canonical 1,800 trial records. Pooled values correspond to the primary intent-family analysis; generator-specific columns show the consistency of the profile across generators.

intent family	Pooled	GPT-Image-2	Qwen-Image-2.0
Concrete control	9.27	9.57	8.98
Visual ambiguity	7.12	7.54	6.70
Narrative	6.82	7.50	6.15
Counterfactual absence	6.75	7.25	6.26
Abstract relation	5.55	7.17	3.92
Layered emotion	4.54	5.79	3.29
Intentionality	3.47	4.40	2.54
Overall	6.07	6.91	5.23

GPT-Image-2 is higher overall by 1.68 points. The intent-family profiles are highly similar across generators ($\rho_{\text{rank}} = 0.964$): both place concrete control first and abstract relation, layered emotion, and intentionality in the bottom three; narrative and counterfactual absence exchange positions. Intent-level cross-generator rank agreement is $\rho = 0.650$.

C.2 Target-Conditioned Metrics and Candidate Selection

CLIP similarity and the official VQAScore implementation are joined to the same 1,800 image records used for PAIR scoring. At image level ($n = 1{,}800$), PAIR is moderately correlated with the comparison selectors: Pearson $r = 0.45$ with CLIP, $r = 0.54$ with VQAScore, and $r = 0.42$ with the target-aware VLM (Spearman $\rho = 0.45$, 0.57, and 0.42, respectively). Thus, the selectors provide related but non-redundant ranking signals, consistent with the candidate-selection disagreements reported below. VQAScore follows the prompt-conditioned formulation of Lin et al. (2024): it uses the generated image and target text to estimate the likelihood of a positive answer. It provides a strong prompt-conditioned comparator to the observer-side PAIR endpoint.

Candidate disagreement is computed over the six images for each intent. PAIR maxima tie in 45.7% of pools and contain 2.04 candidates on average. For two metrics with top sets A and B, the expected probability of agreement under uniform tie resolution is $|A \cap B|/(|A||B|)$; disagreement is one minus this quantity. Averaging this value over intents yields 85.1% disagreement for PAIR–CLIP, 72.2% for PAIR–VQAScore, and 78.9% for PAIR–target-aware VLM. This statistic measures disagreement between the candidate selections induced by each pair of selectors while accounting for tied maxima.

C.3 Rubric-Swap Specificity

For each intent, one recorded prompt-agnostic description from the no-calibration condition is scored three ways: with its correct rubric, with a rubric drawn from another intent in the same intent family, and with a rubric drawn from a different coarse sampling group. For this diagnostic only, the seven families are grouped as {concrete control}, {visual ambiguity, narrative}, {counterfactual absence, abstract relation}, and {layered emotion, intentionality}. These groups are used only to construct the rubric-swap control. Swap assignments use a fixed random seed. Correct, same-family, and cross-group means are 7.19, 0.73, and 0.37.

C.4 Target-Aware and Same-Generator Selection

The target-aware VLM selector is a matched control for PAIR: it uses the same describing model, prespecified rubrics, DeepSeek scorer, score scale, and aggregation, but exposes the verbalized intent when the image description is produced. Each target-aware description is scored signal by signal against the identical rubric, and the selector score averages three evaluations. Thus, the comparison holds the rubric-based scoring pipeline fixed while varying target access during observation.

Like CLIP and VQAScore, this score is used only to select a candidate, and every selected candidate is evaluated on the identical held-out PAIR endpoint. Table 5 reports the pooled 300-pool analysis alongside the same-generator variant over 600 intent–generator three-candidate pools. The Random row selects uniformly within each pool. PAIR-based selection achieves the highest held-out endpoint score.

Table 5: Selection regret under the identical PAIR-score endpoint. The pooled analysis covers the 300 six-candidate pools; the same-generator analysis covers 600 intent–generator three-candidate pools. Intervals are deterministic 95% bootstrap intervals over intents.

Analysis	Selector	Selected mean	Regret	95% CI
Pooled	PAIR-based	7.50	—	—
Pooled	CLIP	6.41	1.09	[0.89, 1.30]
Pooled	VQAScore	6.85	0.65	[0.46, 0.85]
Pooled	Target-aware VLM	6.52	0.97	[0.83, 1.13]
Pooled	Random	6.07	1.43	[1.27, 1.59]
Same-generator	PAIR-based	6.67	—	—
Same-generator	CLIP	6.19	0.48	[0.38, 0.59]
Same-generator	VQAScore	6.39	0.28	[0.19, 0.37]
Same-generator	Target-aware VLM	6.27	0.40	[0.32, 0.48]
Same-generator	Random	6.07	0.60	[0.52, 0.69]

Table 6 stratifies the same-generator analysis by generator and intent family. Regret is positive in 55 of the 56 cells—the one exception is VQAScore on GPT-Image-2 narrative pools (-0.05)—and it is generally larger in lower-scoring intent families, where PAIR scores leave more headroom for a target-conditioned selector to lose. Intervals widen in the 30-pool concrete-control cells and are noted per cell; the estimator is the same split-sample cross-fitting as Table 5.

Table 6: Split-sample selection regret by generator and intent family. Each cell pools one generator's three candidates per intent over the intent family's intents (n pools); intervals are deterministic 95% bootstrap intervals over intent pools. Concrete-control cells have 30 pools, all others 45.

intent family	n	CLIP	VQAScore	Target-aware	Random
GPT-Image-2					
Concrete control	30	0.20 [−0.02,0.50]	0.10 [−0.12,0.38]	0.14 [0.02,0.31]	0.14 [0.02,0.31]
Visual ambiguity	45	0.40 [0.08,0.77]	0.06 [−0.13,0.26]	0.22 [0.01,0.44]	0.20 [0.01,0.40]
Narrative	45	0.13 [−0.12,0.40]	−0.05 [−0.27,0.19]	0.06 [−0.14,0.26]	0.15 [−0.03,0.33]
Counterfactual absence	45	0.26 [−0.07,0.66]	0.38 [−0.00,0.79]	0.32 [0.00,0.69]	0.50 [0.23,0.81]
Abstract relation	45	0.59 [0.26,0.95]	0.20 [−0.15,0.57]	0.53 [0.19,0.89]	0.66 [0.36,0.97]
Layered emotion	45	0.80 [0.32,1.31]	0.27 [−0.11,0.69]	0.43 [0.07,0.77]	0.60 [0.27,0.94]
Intentionality	45	0.75 [0.37,1.19]	0.52 [0.18,0.89]	0.57 [0.30,0.86]	0.70 [0.43,0.99]
Qwen-Image-2.0					
Concrete control	30	0.37 [0.12,0.65]	0.30 [0.07,0.58]	0.25 [0.10,0.40]	0.49 [0.30,0.69]
Visual ambiguity	45	0.19 [−0.12,0.51]	0.31 [−0.02,0.64]	0.35 [0.09,0.62]	0.56 [0.27,0.84]
Narrative	45	0.48 [0.11,0.88]	0.11 [−0.17,0.41]	0.14 [−0.09,0.38]	0.45 [0.14,0.79]
Counterfactual absence	45	0.62 [0.19,1.15]	0.25 [−0.04,0.58]	0.58 [0.25,0.97]	0.93 [0.62,1.29]
Abstract relation	45	0.53 [0.12,0.97]	0.29 [−0.09,0.72]	0.71 [0.35,1.08]	1.42 [0.93,1.93]
Layered emotion	45	0.81 [0.44,1.22]	0.69 [0.29,1.15]	0.70 [0.41,1.00]	0.80 [0.55,1.07]
Intentionality	45	0.48 [0.15,0.87]	0.46 [0.12,0.83]	0.52 [0.24,0.83]	0.64 [0.36,0.95]

C.5 Open-Weight Observer Replication

An open-weight observer, Qwen3-VL-32B-Instruct served through an OpenAI-compatible relay, re-observed all 1,800 benchmark images under the identical prompt-agnostic observer prompt, and the same DeepSeek-V4-Pro scorer, rubrics, and normalization were used to assign scores. The run completed on 2026-08-01 and covers the full $300 \times 2 \times 3$ trial set. For each record key, we retain the first successful observation.

Intent-level agreement is high against both closed observers: Spearman $\rho = 0.920$ and Pearson $r = 0.924$ against GPT-5.5 over 300 intents, and $\rho = 0.921$ and $r = 0.920$ against Gemini-3.5-Flash. As in the closed-observer transfer of Appendix A, the finer trial-level agreement is more moderate: within-intent Pearson averages $r = 0.50$ (263 intents) against the primary observer and $r = 0.52$ (265 intents) against the secondary. Overall means are 6.12 for the open observer, 6.07 for GPT-5.5, and 6.53 for Gemini-3.5-Flash.

The open-weight observer reproduces the complete intent-family ordering: concrete control 9.36 [8.99, 9.65], visual ambiguity 7.06 [6.41, 7.67], narrative 6.88 [6.32, 7.42], counterfactual absence 6.54 [5.92, 7.16], abstract relation 5.79 [5.26, 6.33], layered emotion 4.88 [4.30, 5.49], and intentionality 3.40 [2.75, 4.09] (95% bootstrap intervals over intents).

Table 7 reports selection transfer in both directions. In the pooled analysis, gains over CLIP and VQAScore are positive with intervals excluding zero whether the open observer evaluates GPT-5.5 selections or selects for GPT-5.5 evaluation. In same-generator pools, gains over CLIP remain positive with intervals excluding zero, while the smaller VQAScore estimates have intervals that include zero.

The open-observer score distribution has substantial endpoint concentration: 448 of 1,800 scores (24.9%) equal 10.0, concentrated mainly among concrete-control items, while 195 (10.8%) equal 0.0, concentrated mainly among the lower-scoring intent families.

Table 7: Cross-observer selection gains with the open-weight observer as one side. The selecting observer's PAIR scores choose a candidate; the evaluating observer's PAIR scores, with the scorer and rubrics unchanged, grade it. Pooled rows cover the 300 six-candidate pools; same-generator rows cover 600 intent–generator three-candidate pools. Intervals are deterministic 95% bootstrap intervals over intents.

Pools	Selects	Evaluates	Gain over CLIP	Gain over VQAScore
Pooled (300)	GPT-5.5	open-weight	0.62 [0.35, 0.88]	0.45 [0.19, 0.70]
Pooled (300)	open-weight	GPT-5.5	0.66 [0.45, 0.87]	0.22 [0.02, 0.43]
Same-generator (600)	GPT-5.5	open-weight	0.22 [0.08, 0.36]	0.09 [−0.04, 0.22]
Same-generator (600)	open-weight	GPT-5.5	0.22 [0.12, 0.32]	0.02 [−0.07, 0.11]

C.6 Open-Weight Scorer Replication

Agreement with the primary DeepSeek-V4-Pro scorer is high: image-level Spearman $\rho = 0.896$ and intent-level $\rho = 0.935$. The intent-family profiles are also highly correlated (Spearman $\rho = 0.964$), although their exact ordering is not identical. Overall means are 6.07 for DeepSeek and 6.65 for the open scorer, indicating a systematic calibration shift, with the largest offset on layered emotion.

Table 8 summarizes the intent-family results. Concrete control remains the highest-scoring intent family under both scorers, while abstract relation, layered emotion, and intentionality remain the bottom three.

Selection regret also remains positive under the open scorer. Table 9 reports regret relative to PAIR-based selection for CLIP, VQAScore, and the target-aware VLM selector. All three selectors retain positive mean regret with 95% confidence intervals excluding zero under both scorers. The relative ordering among the three selectors is not identical, so we treat this result as robustness of the positive-regret finding rather than selector-rank stability.

Table 8: PAIR scores by intent family under DeepSeek and Qwen3-32B on 1,800 images, averaging scores over the same three observer descriptions per image. Diff. denotes DeepSeek − Qwen3.

Intent family	DeepSeek	Qwen3	Diff.
Concrete control	9.27	9.11	+0.17
Visual ambiguity	7.12	7.49	−0.37
Narrative	6.82	7.50	−0.68
Counterfactual absence	6.75	7.15	−0.39
Abstract relation	5.55	6.36	−0.82
Layered emotion	4.54	5.87	−1.33
Intentionality	3.47	3.88	−0.42

Table 9: Selection regret relative to PAIR-based selection under the two scorers. Intervals are 95% bootstrap CIs over 300 intent pools.

Scorer	Selector	Regret	95% CI
DeepSeek	CLIP	1.09	[0.89,1.30]
	VQAScore	0.65	[0.46,0.85]
	Target-aware	0.97	[0.83,1.13]
Qwen3	CLIP	0.86	[0.67,1.06]
	VQAScore	0.53	[0.36,0.72]
	Target-aware	0.89	[0.75,1.04]

C.7 Intent-Family Covariate Model

We evaluate the intent-family profile after accounting for prompt length, rubric size, and generator identity. We fit an ordinary-least-squares model on the 600 intent–generator mean PAIR scores with intent-family indicators (reference: abstract relation), standardized prompt word count, standardized rubric signal count, and a generator indicator, using HC1 heteroskedasticity-consistent standard errors. Prompt length and rubric size have small estimated associations: prompt word count -0.14 (SE $= 0.11$) and rubric signal count 0.10 (SE $= 0.10$) per standard deviation, while the generator indicator is -1.68 (SE $= 0.19$) for Qwen-Image-2.0 relative to GPT-Image-2. The adjusted intent-family means in Table 10 preserve the descriptive ordering from Figure 3.

C.8 Failure-Attribution Case Studies

Table 11 chains six representative trials from intent to image to prompt-agnostic description to signal-level scoring, drawn verbatim from the archived trial records. The PAIR score averages the three

Table 10: Adjusted intent-family means from the covariate model (600 intent–generator rows), evaluated at mean covariate values and averaged equally across the two generators. The ordering matches the unadjusted profile.

intent family	Adjusted mean
Concrete control	9.05
Visual ambiguity	7.11
Narrative	6.74
Counterfactual absence	6.72
Abstract relation	5.56
Layered emotion	4.65
Intentionality	3.63

observer/scorer calls; the target-aware column is the same image's target-aware VLM score (the target-aware selector of the regret analysis), which sees the image together with the intent. The contrast between the two columns provides a diagnostic view of three recurrent patterns: (a) both prompt-agnostic and target-aware readings assign little evidence to the image; (b) the target-aware reading identifies evidence that is absent from the prompt-agnostic description; and (c) the prompt-agnostic description recovers the target evidence only partially. The `intent_01` row is the trial displayed in Figure 1.

These cases illustrate how low PAIR can arise from different observable patterns. In `intent_01` and `abst_15`, the target-aware VLM assigns high scores to the same image while the prompt-agnostic description omits the decisive intent-specific evidence. In `intent_10` and `abst_06`, both scores are low, consistent with weaker support for the target in the image itself. The `emot_04` case lies between these extremes: the observer identifies departure or transition but remains uncertain about the intended reluctant emotional state. The success case `intent_09` shows that even an intentionality item can receive high PAIR when the image provides a distinctive cue.

D Human Study Details

Three main participant-facing human studies and two supplementary checks involve 249 participants and roughly 5,000 judgments and descriptions in total. The main studies examine open-ended prompt-agnostic descriptions, prompt-agnostic intent identification, and intent-informed image preference. A separate five-rater annotation study validates the target-aware scorer in Appendix D.4.

D.1 Open-Ended Recovery Protocol and Per-Intent Accounting

The open-ended study recruited 49 participants (686 open-ended descriptions) for a Chinese-language image-description task covering 28 stratified intents drawn from the seven benchmark intent families. The intents were split into two complementary 14-intent forms (an A/B split), so each participant described one half of the set. Participants saw each image without the intent and wrote an open-ended description under a matched Chinese translation of the VLM observer instruction: report what the image shows and what it seems to convey from the image alone. Every human description was then scored by the same target-aware scorer and the same pre-specified rubrics used for the VLM descriptions, yielding one human PAIR score per description. Participants wrote in Chinese, and the multilingual scorer applied the English rubrics directly to the Chinese descriptions without an intermediate machine-translation step. Table 12 reports the per-intent accounting; across the 28 intents the human means and VLM PAIR scores correlate at Spearman $\rho = 0.715$ ($p = 1.9 \times 10^{-5}$) and Pearson $r = 0.751$. To evaluate correspondence beyond between-family differences, we mean-center both series within each intent family. The resulting correlation remains essentially unchanged (Spearman $\rho = 0.720$, $p = 1.6 \times 10^{-5}$; Pearson $r = 0.745$), so the agreement also holds within intent families.

Table 11: Representative success and failure chains from archived trial records. Observer cells quote the decisive fragment of the first prompt-agnostic description. Signal scores are out of 2. PAIR averages three observer/scorer calls; target-aware is the target-aware VLM score for the same image.

Image	Intent / record	prompt-agnostic description	Signal scores	Outcome
	A person pretending to read while actually eavesdropping on the conversation nearby `intent_09`; intentionality; GPT-Image-2, t0	"holding a dark blue hardcover book up so only his eyes … show. He looks sideways over the top of the book toward … a woman and a man … talking closely"	pretending to read: 2 eavesdropping: 2	PAIR 10.0 Target 10.0 **Success**
	A person knocking over a cup on purpose while pretending it was an accident `intent_01`; intentionality; GPT-Image-2, t1	"caught mid-spill, tipping a mug … His wide eyes and open mouth show surprise or alarm … a sudden mishap or clumsy moment"	knocking over: 2 intentional: 0 pretending accident: 0	PAIR 3.3 Target 10.0 **Target-aware high**
	A person who keeps taking advantage of another who gets nothing in return `abst_15`; abstract relation; Qwen-Image-2.0, t1	"resting his hand on the shoulder of a younger woman … suggests a relationship of authority, support, or possibly concern, though the exact nature is unclear"	exploitative transfer: 0 no reciprocation: 0 repeated/ongoing: 0	PAIR 0.0 Target 10.0 **Target-aware high**
	A person about to leave a long-time home, calm on the surface, full of reluctance within `emot_04`; layered emotion; GPT-Image-2, t0	"leaving home or arriving after a trip, paused at the threshold … a moment of departure, transition, or return"	three calls: 1/1/1 1/0/0 1/0/1	PAIR 3.3 Target 10.0 **Partial**
	A person turning a photo face-down, not wanting anyone to see it `intent_10`; intentionality; Qwen-Image-2.0, t0	"looking intently at a couple of photographs or prints he is holding … private and introspective"	turning action: 0 face-down state: 0 concealment intent: 0	PAIR 0.0 Target 2.2 **Both low**
	A person utterly marginalized within a group, with almost no presence `abst_06`; abstract relation; Qwen-Image-2.0, t1	"one man in a gray hoodie in the center looking down and apart … familiarity, camaraderie, and upbeat energy"	spatial marginalization: 0 visual diminishment: 0 social disconnection: 0	PAIR 0.0 Target 3.3 **Both low**

D.2 Pairwise Study: Items and Response Accounting

The pairwise study uses a 20-item evaluation set finalized before human data collection. Each item displays an intent and two anonymized images with randomized left/right order: a PAIR-guided refined output from the structured observer-feedback pipeline and a repeated-uncalibrated-generation comparator. The four response options are:

1. the first image communicates the intent better;
2. the second image communicates the intent better;
3. no clear difference between the two images;
4. neither image communicates the intent well.

The private export contained 50 rows. Under the predefined audit rule, both rows in one same-display-name duplicate group were excluded, while one row was retained from a separate same-display-name exact-response duplicate pair. This leaves 47 participants in the released 20-column

Table 12: Per-intent accounting for the open-ended human study. Human values average 23 or 26 descriptions per intent; VLM PAIR is the benchmark score for the same image. Across the 28 intents, $\rho = 0.715$ and $r = 0.751$.

Intent	Human mean	Human SD	n	VLM PAIR
absent_13	6.11	3.00	26	7.80
absent_26	2.93	2.92	23	10.00
absent_42	6.04	1.73	26	7.50
absent_44	2.83	2.33	23	0.00
abst_26	6.02	2.73	26	8.30
abst_28	8.47	1.07	26	8.30
abst_36	0.69	1.23	23	1.10
abst_45	8.77	2.50	23	10.00
ambig_37	6.60	1.32	23	10.00
ambig_39	5.85	2.60	26	7.50
ambig_40	2.97	1.13	23	2.20
ambig_41	5.10	1.25	26	7.50
conc_06	8.10	1.30	23	10.00
conc_07	7.43	1.62	26	10.00
conc_21	8.50	1.78	23	6.70
conc_28	6.54	2.95	26	10.00
emot_02	1.90	1.92	26	6.10
emot_35	4.53	1.62	23	2.80
emot_40	5.83	2.82	23	10.00
emot_41	3.93	1.43	26	5.80
intent_09	8.32	2.32	23	10.00
intent_10	1.37	0.67	23	0.00
intent_14	3.03	0.93	26	3.30
intent_37	3.81	1.40	26	3.30
narr_06	6.90	2.18	26	7.80
narr_16	7.37	2.09	26	7.80
narr_18	5.47	1.29	23	3.30
narr_31	9.57	2.09	23	10.00

matrix and $47 \times 20 = 940$ judgments. The hidden item key maps first/second responses to the PAIR-guided refined image. The last two options are *non-decisive choices*; they are not participant exclusions. Figure 7 shows three items in the form participants saw them.

The all-judgment target rate is $573/940 = 61.0\%$. The decisive rate is $573/757 = 75.7\%$ after the 183 non-decisive choices, not respondents, are omitted from that denominator. The rates describe this 20-item evaluation set. The per-participant decisive win rate averages 0.766 (median 0.737) and exceeds chance (one-sample $t(46) = 11.39$, $p = 5.5 \times 10^{-15}$; Wilcoxon $p = 9.2 \times 10^{-9}$). Participant-cluster bootstrap 95% CIs (10,000 resamples) are [0.56, 0.66] for the all-judgment rate and [0.71, 0.80] for the decisive rate. Across items, the decisive win rate ranges from 14/34 (AD11) to 42/45 (AD1).

The 20 items and their image pairs were finalized before any response was collected; no item was added, dropped, or re-edited afterward. AD1 provides an informative ceiling case: both images receive a PAIR score of 10.0, yet participants select the PAIR-guided refined image in 42 of 45 decisive judgments. This case shows that preference and composition can continue to distinguish images after intent recovery reaches ceiling.

The landing page described the task as academic research and instructed respondents not to use AI tools. The private operational export included self-entered display names for duplicate review; the released matrix removes names, response IDs, timestamps, duration, and other free-text metadata. No demographic fields were collected. Participants received a small payment of approximately 10 RMB, as in the open-ended study. The released `anonymized_responses.csv`, `human_study_protocol.json`, and summary builder reproduce every count in Table 13; the raw export remains private.

Figure 7: **Example items from the intent-informed pairwise study.** Participants saw the intended meaning and two anonymized images in randomized order: a PAIR-guided refined output and a repeated-uncalibrated comparator. Positions are normalized here with the refined output on the left; values below images are their recorded PAIR scores. Fractions next to item identifiers give PAIR-guided choices over decisive judgments (Table 13).

AD1 (42/45)
"A shop's rolling shutter slowly coming down as the last customer leaves"



PAIR-guided (PAIR 10.0)



repeated uncalibrated (PAIR 10.0)

AD15 (43/47)
"A person assimilated under group pressure, giving up their own stance"



PAIR-guided (PAIR 9.2)



repeated uncalibrated (PAIR 5.0)

AD11 (14/34)
"A person asking a question just as the other is about to leave, deliberately holding them back"



PAIR-guided (PAIR 8.9)



repeated uncalibrated (PAIR 3.3)

D.3 PAIR Selection Validation: Protocol and Per-Pair Accounting

This study asks whether the selector disagreements quantified in Section 5 are perceptible to prompt-agnostic human observers. Without manual image screening, we retained pools where PAIR and an alignment metric selected different images, ranked them by held-out PAIR-score gap, and selected four pairs per intent family—two against CLIP and two against VQAScore. This yields 28 pairs; all gaps are at least 2.5 points (mean 5.8), and the item list was fixed before human responses.

Each trial shows one image without the prompt and two interpretations: the true intent and a foil derived from the recorded prompt-agnostic reading of the comparison-arm image. Qwen3-32B (temperature 0) only normalizes the foil to match the correct option in style, abstraction level, and length. Both image arms use the identical option pair, with arm assignment and option order counterbalanced across two forms; each participant sees exactly one arm per pair. Because the foil originates

Table 13: Response accounting for the 20 intent-informed pairwise items. "Target" denotes the PAIR-guided refined image after applying the hidden left/right key. Non-decisive combines "no clear difference" and "neither." Every row totals 47.

Item	Target choices	Decisive choices	Non-decisive choices
AD1	42	45	2
AD2	26	45	2
AD3	34	39	8
AD4	38	42	5
AD5	34	40	7
AD6	31	45	2
AD7	33	40	7
AD8	36	43	4
AD9	28	36	11
AD10	20	33	14
AD11	14	34	13
AD12	29	38	9
AD13	27	30	17
AD14	17	31	16
AD15	43	47	0
AD16	24	40	7
AD17	26	29	18
AD18	22	37	10
AD19	26	32	15
AD20	23	31	16
Total	**573**	**757**	**183**

from the comparison-arm image, its plausibility need not be matched across arms, so this experiment is interpreted as a diagnostic of rule-selected disagreements rather than a benchmark-wide effect. The open-ended recovery study provides complementary grounding without forced-choice distractors.

Across 53 participants and 1,484 valid judgments, PAIR-selected images yield 85.9% identification accuracy versus 39.4% for alignment-selected images, a $+46.6$-point difference (pair-cluster bootstrap 95% CI [$+36.2$, $+56.8$]). PAIR-selected images lead in 26 of 28 pairs, with similar gaps against CLIP ($+46.8$) and VQAScore ($+46.3$). Per-pair results appear in Table 14.

Because judgments are cross-classified by participant and pair, we additionally compute a two-way cluster bootstrap resampling participants and pairs jointly (10,000 resamples); the pooled gap retains a 95% CI of [$+34.1$, $+59.0$] points, consistent with the pair-cluster interval above.

The landing page, consent, payment, and de-identification followed the same protocol as the pairwise study; the released `anonymized_responses.jsonl`, `survey.json`, and summary builder reproduce every count in Table 14.

D.4 Human Validation of the Target-Aware Scorer

We independently validate the primary target-aware scorer against five bilingual human raters using the 28 stratified intents from the open-ended study. For each intent, we randomly sample three previously collected human descriptions and use the first recorded VLM-observer description, yielding 112 descriptions and 352 description–signal items. The item set was fixed before annotation. Raters see only the observer description, one pre-specified rubric signal, and its YES/PARTIAL/NO criteria; they do not see the image, full intent, automated score, or provenance label.

Human ratings use the rubric-native 0/1/2 scale. For categorical agreement, the automated scorer's rare half-step outputs (6/352 items) are mapped to PARTIAL; retaining the original five-level outputs gives essentially unchanged agreement ($\kappa = 0.877$ vs. 0.868). Confidence intervals use 10,000 cluster-bootstrap resamples of the 28 intents.

Table 14: Per-pair prompt-agnostic intent-identification accuracy in the PAIR-based selection validation study. Each row is a rule-selected selector-disagreement case. Accuracy is the share of uninformed participants choosing the intended interpretation for the PAIR-selected or alignment-selected image; n gives valid judgments per arm (PAIR/metric).

Pair	Metric	Intent family	Gap	PAIR-selected (%)	Metric-selected (%)	n
HS01	CLIP	concrete control	2.8	96	24	28/25
HS02	CLIP	concrete control	2.7	88	89	25/28
HS03	VQAScore	concrete control	3.3	96	96	28/25
HS04	VQAScore	concrete control	2.5	80	14	25/28
HS05	CLIP	visual ambiguity	3.3	100	32	28/25
HS06	CLIP	visual ambiguity	3.3	100	57	25/28
HS07	VQAScore	visual ambiguity	3.3	96	84	28/25
HS08	VQAScore	visual ambiguity	2.7	100	57	25/28
HS09	CLIP	narrative	5.0	32	16	28/25
HS10	CLIP	narrative	4.1	84	25	25/28
HS11	VQAScore	narrative	5.6	79	52	28/25
HS12	VQAScore	narrative	2.8	100	29	25/28
HS13	CLIP	counterfactual absence	9.4	93	0	28/25
HS14	CLIP	counterfactual absence	5.8	44	46	25/28
HS15	VQAScore	counterfactual absence	6.7	93	32	28/25
HS16	VQAScore	counterfactual absence	6.6	92	18	25/28
HS17	CLIP	abstract relation	9.1	82	28	28/25
HS18	CLIP	abstract relation	8.3	100	54	25/28
HS19	VQAScore	abstract relation	10.0	86	16	28/25
HS20	VQAScore	abstract relation	9.4	100	7	25/28
HS21	CLIP	layered emotion	8.3	89	56	28/25
HS22	CLIP	layered emotion	6.6	92	21	25/28
HS23	VQAScore	layered emotion	7.2	93	72	28/25
HS24	VQAScore	layered emotion	5.6	56	50	25/28
HS25	CLIP	intentionality	6.7	100	32	28/25
HS26	CLIP	intentionality	6.7	92	57	25/28
HS27	VQAScore	intentionality	10.0	86	12	28/25
HS28	VQAScore	intentionality	4.4	56	25	25/28

Table 15 summarizes the results. Scorer–human consensus agreement is high at the signal level ($\kappa = 0.868$ [0.818, 0.907], exact agreement 83.0%, MAE 0.18), and human- and model–scored description–level PAIR agree closely ($\rho = 0.911$ [0.848, 0.943], $r = 0.908$, MAE 0.82). Agreement with individual raters ($\kappa = 0.743$–0.830) is comparable to individual-rater agreement with the remaining-human consensus (0.730–0.793). Results remain strong within both observed source/language strata (Chinese human descriptions: $\kappa = 0.891$; English VLM descriptions: $\kappa = 0.794$); source and language are confounded by design.

Level	Comparison	n	Agreement	Error
Signal	Human–human	352	$\alpha = 0.70$; pairwise $\kappa = 0.65$–0.77	—
Signal	Scorer–human consensus	352	$\kappa = 0.87$ [0.82, 0.91]; exact $= 0.83$	MAE $= 0.18$
Description	Scorer PAIR–human PAIR	112	$\rho = 0.91$ [0.85, 0.94]; $r = 0.91$	MAE $= 0.82$
Signal	Chinese human descriptions	264	$\kappa = 0.89$ [0.85, 0.93]; exact $= 0.84$	MAE $= 0.16$
Signal	English VLM descriptions	88	$\kappa = 0.79$ [0.65, 0.90]; exact $= 0.80$	MAE $= 0.24$

Table 15: Human validation of the target-aware scorer. Human–human exact agreement requires all five raters to coincide (46.6%). Bracketed intervals are 95% cluster bootstraps over the 28 intents.

D.5 Supplementary Human Checks

Two supplementary studies further characterize the relationship between human and VLM recovery. In a four-way forced-choice study on the same 28 intents, 61 participants recover the exact intent at 36% across 1,708 judgments (chance: 25%). Three VLM observer families achieve 33–55% across 401 valid judgments: GPT-5.5 reaches 40%, Gemini-3.5-Flash 55%, and Qwen3-VL-32B 33%. Both human and VLM recovery are above chance and remain well below ceiling.

In a separate informed-rating study ($N = 39$), ratings average 6.84 when the intent is visible and 5.53 under prompt-agnostic evaluation, a between-group difference of approximately 1.3 points. Together, these studies show comparable human and VLM difficulty in exact intent identification and quantify the effect of revealing the target during evaluation.

E PAIR-Guided Refinement: Pipeline and Evaluation Details

E.1 Pipeline Mechanics and Data Mapping

All four conditions start from one initial generated image and may produce up to five additional images, so every condition has the same maximum generation budget of six images per intent. Conditions A and B stop early when PAIR reaches 9; conditions C and D always produce all five continuation outputs.

Structured observer feedback (A).
The current image is described without target access and scored against the target rubric. The planner receives the prompt-agnostic observer description and scalar PAIR score; it has access to the target and rubric and uses them to diagnose missing signals before editing or regenerating the image. Each evaluated round, including the terminal output, averages three observer/scorer calls. The loop stops when PAIR reaches 9 or after at most five refinement rounds, and the terminal evaluated output is reported rather than the best historical round.

Self-guided refinement (B).
The same generator/planner family inspects the image with target and rubric access and plans the next edit or regeneration without an external prompt-agnostic description. Its terminal PAIR score averages three observer/scorer calls.

Text-only optimization (C).
Five independent text-only prompt rewrites start from the original intent, each producing one image without visual feedback. Each image receives one observer/scorer call; the five scores are then averaged.

No calibration (D).
The record stores the original image and five continuations generated independently from the original prompt without feedback. Each continuation receives one observer/scorer call. The current reporting mean averages only those five continuation scores and excludes the stored original image.

The four conditions represent complete pipeline configurations. Conditions A and B provide the closest comparison of refinement with and without an external prompt-agnostic observer, while C and D provide non-adaptive references with different evaluation aggregation. Figure 8 displays three worked A-versus-D example pairs drawn from these archived records.

E.2 Aggregate Tests



Figure 8: **Qualitative examples from the PAIR-guided refinement probe.** Each intent compares the final output of structured observer feedback (A) with one displayed continuation from the no-calibration condition (D). Values below the images are recorded image-level PAIR scores. Condition D contains five non-adaptive continuations per intent; the displayed continuation is illustrative and need not equal the condition-level mean. Aggregate results over all 300 intents are reported in Table 16.

Table 16: Aggregate PAIR scores for the four refinement pipelines. A and B report terminal adaptive outputs averaged over three evaluation calls; C and D average five single-evaluation outputs.

Pipeline (record key)	Mean	SD	Reporting unit
Structured observer feedback (A)	8.77	2.10	final adaptive output
Self-guided refinement (B)	7.98	2.31	final adaptive output
Text-only optimization (C)	6.66	2.25	mean of five rewrites
No calibration (D)	7.12	2.31	mean of five continuations

The headline contrast pairs each intent's structured-feedback final output against its no-calibration continuation mean: a mean difference of $+1.65$ (95% CI [1.41, 1.89]), paired $d_z = 0.78$, $t(299) = 13.48$, $p = 1.1 \times 10^{-32}$, over 300 intents.

E.3 Cross-Evaluator Replication

We evaluate whether the refinement gap transfers to an independent open-weight stack. All 300 condition-A terminal images are re-observed by Qwen3-VL-32B-Instruct with three calls each, and all 1,500 condition-D continuations are re-observed once each. Qwen3-32B scores every resulting description against the same pre-specified rubrics.

Table 17: Cross-evaluator replication of the refinement gap over 300 intents. Rows pair each condition's observer and scorer; the A−D gap contrasts the row pair marked with the same symbol. The open-stack rows use identical observer and scorer families for both conditions, so the comparison is symmetric.

Pair	Observer	Scorer	A	D
†	GPT-5.5 (primary)	DeepSeek (primary)	8.77	7.12
‡	Qwen3-VL-32B	Qwen3-32B	7.97	7.28
§	Qwen3-VL-32B (A), GPT-5.5 (D)	DeepSeek (primary)	7.61	7.12

A−D gap	Mean	95% CI	Test
† primary stack	$+1.65$	[1.41, 1.89]	$t = 13.5$, $p = 1.1 \times 10^{-32}$
‡ open stack, symmetric	$+0.69$	[0.45, 0.93]	$t = 5.7$, $p = 2.2 \times 10^{-8}$
§ mixed, open-observer A	$+0.44$	[0.13, 0.74][a]	$t = 2.9$

[a] Asymmetric pairing: the open-observer A mean is contrasted with the primary-stack D mean under the shared primary scorer.

We report two separate diagnostics to contextualize the attenuation from $+1.65$ to $+0.69$. Replacing the GPT-5.5 observer with Qwen3-VL lowers the A mean from 8.77 to 7.61 (-1.16) under the shared primary scorer: the open observer's prompt-agnostic descriptions cover fewer rubric signals, consistent with the lower PAIR scores produced by the open observer on this evaluation. Holding the description texts fixed, the open scorer is mildly more lenient than the primary scorer ($+0.37$ on the open-observer A texts; $+0.61$ on the primary-stack D texts), and the score increase appears in both conditions. Under the fully open, symmetric stack, condition A retains a mean advantage of $+0.69$ points (95% CI [0.45, 0.93], paired $t(299) = 5.75$, $p = 2.2 \times 10^{-8}$, $d_z = 0.33$). The largest gains occur in abstract relation ($+1.04$), counterfactual absence ($+1.10$), layered emotion ($+0.93$), and narrative ($+0.80$), while concrete control is at ceiling in both conditions (-0.02).

F Auditability and Released Artifacts

PAIR records preserve the intermediate observer descriptions and signal-level scores, allowing aggregate results to be traced to individual image–description pairs. Absolute scores may vary with the observer, scorer, service version, language, and prompt configuration; the released records therefore include the model information, timestamps, prompts, rubrics, and per-call outputs used in the reported experiments.

Generated images in Figures 1 and 4 are tied to exact trial records. The reproducibility manifest stores the source-family tag, intent, model, condition, trial index, source path, stable paper path, and SHA-256 digest for every displayed image. Numerical Figures 3 and 6 are generated from an evidence ledger containing the reported aggregates. Together with the preserved prompts, rubrics, descriptions, per-call scores, model information, and timestamps, these artifacts support reconstruction of the paper's principal results.