A2I-BENCH: MEASURING THE VALUE OF CLARIFICATION IN INTERACTIVE TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive text-to-image systems ask for clarification when a user prompt admits multiple plausible interpretations. Yet ambiguity alone does not determine whether clarification is useful: the same user intent may already match one generator's behavior under the original prompt while conflicting with another's. To formalize this issue, we introduce **A2I-Bench**, a new benchmark of 527 ambiguous text-to-image prompts for measuring the value of clarification **(VoC)**. Each case contains two linguistically plausible, visually distinguishable user intents together with intent-specific visual contracts, enabling matched counterfactual evaluation while holding the prompt, generation system, and image budget fixed. We operationalize clarification as revealing the user's intended interpretation. For each generation system, we estimate a Direct-based default interpretation from independent direct generations and measure the gain from intent revelation relative to a frozen no-question rewrite. Across four generation systems, revealing the intended interpretation improves contract satisfaction by 0.158 when the user's intent conflicts with the Direct-based default, a 0.092 larger gain than when they are aligned. These results show that clarification value depends on whether the no-question generation behavior already realizes the user's intent, rather than on prompt ambiguity alone. Because generator defaults and clarification gains are estimated separately for each system, A2I-Bench provides a reusable protocol for evaluating clarification value in new text-to-image generators. We will release the benchmark and code.

1 INTRODUCTION

Should a text-to-image system ask before it draws? A common answer is to ask when the user prompt is ambiguous. Prior work follows this intuition, using clarification to resolve multiple plausible interpretations of a request before generation (Mehrabi et al., 2023; Hahn et al., 2024). But ambiguity alone does not determine whether asking is useful. What matters is whether the information obtained through clarification changes the generated outcome in a way the user values.

Consider a prompt with two plausible visual interpretations. If a user's intended interpretation already matches what a generator tends to produce, revealing that intent may change little. The same intent, however, may conflict with the default behavior of another generator, making the information substantially more valuable. Moreover, even after the intended interpretation is known, generators can differ in how reliably they realize it. Clarification value therefore depends not only on what is uncertain in the prompt, but also on what the downstream generator would produce before and after that uncertainty is resolved.

This dependence is not captured by treating clarification purely as an information-acquisition problem. Existing work has shown that clarification can improve alignment on ambiguous or underspecified requests (Mehrabi et al., 2023; Hahn et al., 2024), while recent work in language systems studies when additional information is worth the interaction cost (Zhang & Choi, 2025; Sun et al., 2025). Text-to-image generation introduces an additional bottleneck: image generators exhibit different semantic preferences under ambiguous prompts and different capabilities for realizing compositional visual constraints (Cekinmez et al., 2026). The same information about user intent can therefore have different downstream value for different generators.

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

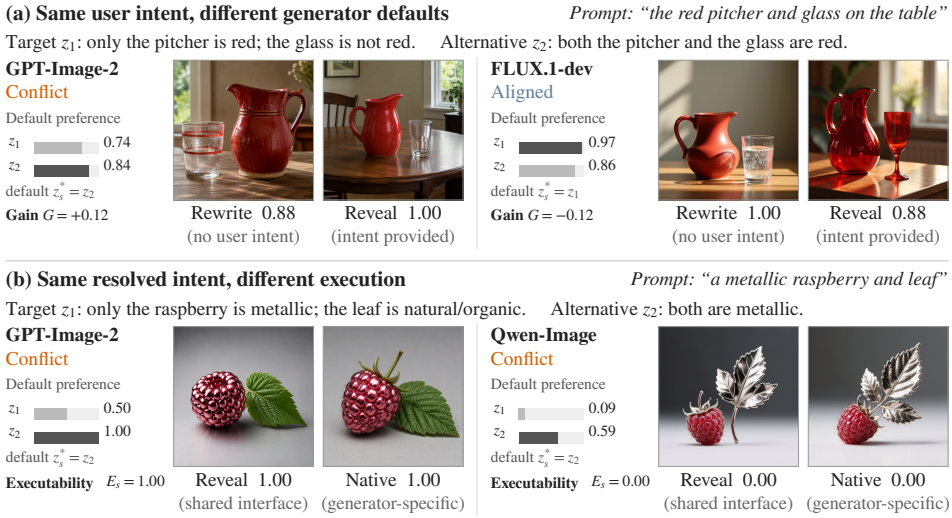


Figure 1: **Clarification value depends on both generator defaults and execution.** For each ambiguous prompt, we compare a no-question route (REWRITE), whose frozen rewriter sees only the original prompt, with an intent-revealed route (REVEAL), which is given the user's intended interpretation. The clarification gain is $G = S_{\text{Reveal}} - S_{\text{Rewrite}}$. **(a)** The same fixed user intent can conflict with one generator but align with another because generators favor different interpretations of the same ambiguous prompt. GPT-Image-2 favors the competing interpretation and benefits from Reveal, whereas FLUX.1-dev already favors the target interpretation and Reveal decreases its contract-satisfaction score in this example. **(b)** Resolving the ambiguity does not guarantee successful rendering. For the same resolved intent, GPT-Image-2 realizes the requested attribute binding, whereas Qwen-Image does not. E_s measures how well the generator renders the resolved intent using its frozen native interface. Scores denote satisfaction of the intent-specific visual contract, i.e., the atomic visual requirements and exclusions defining the intended interpretation.

Measuring this effect requires separating three factors: *prompt ambiguity*, *user intent*, and *generator behavior*. Assigning a single "correct" interpretation to each ambiguous prompt would conflate them. Figure 1 illustrates the distinction: before clarification, different generators may favor different interpretations of the same prompt; after the intended interpretation is revealed, they may still differ in how successfully they render it.

We introduce **A2I-Bench**, a benchmark of 527 ambiguous text-to-image prompts designed to isolate these factors. Each benchmark case defines two linguistically plausible, visually distinguishable user intents for the same prompt. Each intent is represented by an explicit visual contract consisting of atomic requirements and exclusions. We evaluate matched counterfactual user states over the same prompt: one user intends the first interpretation, while another intends the second. The prompt, generation system, and image budget remain fixed; only the intended interpretation changes. This design lets us directly measure how much revealing user intent improves the final visual outcome.

We operationalize clarification as controlled intent revelation. For each generator, we first estimate a Direct-based default interpretation from independent direct generations. We then compare an intent-revealed generation condition against a frozen no-question rewrite that receives only the original prompt. This yields two matched states for each prompt-generator pair: a Direct-aligned state, where the user's intent matches the interpretation preferred under Direct generation, and a Direct-conflict state, where it does not. This independently estimated Direct preference need not exactly match the interpretation realized by the no-question Rewrite route, which may itself resolve part of the ambiguity during reformulation. Importantly, alignment is generator-specific: the same fixed user intent can be aligned with one generator and in conflict with another.

Across 527 A2I-Bench cases and four generation systems, we obtain three findings. First, generators can favor different interpretations of the same ambiguous prompt. Second, clarification gain is larger under Direct-defined conflict: the pooled gain in visual-contract satisfaction is 0.158 in the Direct-conflict state, exceeding the Direct-aligned gain by 0.092. A complementary Rewrite-defined

analysis shows that this contrast more directly tracks the behavior of the no-question route. Third, even with the intent resolved, systems differ in how reliably they realize its visual constraints through their frozen interfaces. Together, these findings show that clarification value depends on both user intent and the downstream generation system.

These observations motivate a generator-conditioned notion of *Value of Clarification* (VoC): the improvement in expected visual satisfaction obtained by revealing the user's intent, optionally discounted by interaction cost. This perspective shifts the question from *"Is the prompt ambiguous?"* to *"Will resolving this ambiguity improve the outcome for this user and this generator?"* Because generator defaults and clarification gains are estimated independently for each system, the same protocol can be applied to new text-to-image generators without assigning a generator-independent "correct" interpretation to the prompt.

Our contributions are threefold:

- We introduce **A2I-Bench**, a benchmark of 527 ambiguous text-to-image prompts that separates prompt ambiguity from user intent through matched counterfactual user states and explicit intent-specific visual contracts.
- We introduce a **generator-conditioned evaluation protocol** that separately measures each system's default interpretation and its ability to realize a revealed intent. Because both are estimated per system, A2I-Bench can be directly applied to new text-to-image generators.
- Across four generation systems, we show that intent revelation provides greater value under Direct-defined conflict, and that this effect is even stronger when alignment is defined by the no-question Rewrite behavior itself. These results show that clarification value depends on the interaction between user intent and the behavior of the downstream generation pipeline rather than on prompt ambiguity alone.

2 Measuring Clarification Value

We study clarification as a counterfactual intervention: for the same prompt, user intent, and generation system, how much does revealing the intended interpretation change the final visual outcome?

2.1 Prompts, intents, and visual satisfaction

Let p be an ambiguous prompt with two plausible visual interpretations,

$$Z(p) = \{z_1, z_2\}. \tag{1}$$

Each interpretation z is associated with a visual contract specifying the properties that should and should not appear in the generated image. We write

$$S(I, z) \in [0, 1] \tag{2}$$

for the contract satisfaction of image I under interpretation z.

A user state specifies which interpretation the user intends. For the primary analysis, we consider specific users with $z_u \in Z(p)$.

2.2 Direct-based interpretation preference

An ambiguous prompt need not be neutral to an image generator. Let s denote a fixed generation system. We characterize its preference under Direct generation, before any information about the user's intended interpretation is provided. Define the Direct support for interpretation z as

$$B_s(p, z) = \mathbb{E}_{I \sim P_s(\cdot|p)} \left[S(I, z) \right]. \tag{3}$$

The system's default interpretation is then

$$z_s^\star(p) = \arg\max_{z \in Z(p)} B_s(p, z). \tag{4}$$

The two fixed user states can be labeled relative to this preference: a user is *Direct-aligned* when $z_u = z_s^\star(p)$, and *Direct-conflict* otherwise. The same fixed user intent can therefore be Direct-aligned with one generation system and Direct-conflict with another.

2.3 CLARIFICATION GAIN

We compare two routes under a matched image budget. In the primary no-question route, REWRITE, a frozen rewriter observes the original prompt p but receives no information about the user's intended interpretation z. In the intent-revealed route, REVEAL, the intended interpretation is supplied explicitly to the generation pipeline.

For a user intending z, define the expected contract satisfaction under the two routes as

$$\begin{aligned} R_s(p, z) &= \mathbb{E}_{I \sim P_s^{\text{Reveal}}(\cdot|p,z)} \left[S(I, z) \right], \\ W_s(p, z) &= \mathbb{E}_{I \sim P_s^{\text{Rewrite}}(\cdot|p)} \left[S(I, z) \right]. \end{aligned} \tag{5}$$

The primary clarification gain is the improvement from revealing the intended interpretation relative to the frozen no-question Rewrite route:

$$G_s(p, z) = R_s(p, z) - W_s(p, z). \tag{6}$$

Our primary question is not whether G_s is positive on average, but whether it differs according to the relation between user intent and the independently estimated Direct preference. We therefore report

$$\nu = \mathbb{E}[G_s(p, z_u) \mid z_u \neq z_s^\star(p)], \tag{7}$$

the mean clarification gain in the Direct-conflict state, and

$$\Delta = \mathbb{E}[G_s(p, z_u) \mid z_u \neq z_s^\star(p)] - \mathbb{E}[G_s(p, z_u) \mid z_u = z_s^\star(p)], \tag{8}$$

the difference in clarification gain between Direct-conflict and Direct-aligned states.

Because the prompt and generation system are fixed across the two counterfactual user states, Δ compares clarification gain according to the user's relation to independently measured Direct behavior. The Direct preference is estimated independently of the REWRITE and REVEAL outcomes; the no-question REWRITE route may nevertheless realize a different interpretation, a distinction examined in Section 4 and Appendix E.5. The Direct-aligned/conflict labels are retrospective diagnostics: they depend on the user's realized intent and are therefore not available to an ask/no-ask policy before clarification. Appendix E.9 separately evaluates prospective decisions using only pre-clarification signals.

2.4 EXECUTION AFTER INTENT REVELATION

Knowing the intended interpretation does not guarantee that a generator can render it. We separately measure resolved-intent executability under the frozen generator-specific NATIVE interface as

$$E_s(p, z) = \mathbb{E}_{I \sim P_s(\cdot|\kappa_s(p,z))} \left[S(I, z) \right], \tag{9}$$

where κ_s denotes the generator-specific NATIVE interface frozen before the scale evaluation.

Thus, B_s characterizes what the system tends to produce under DIRECT generation before intent resolution, while E_s characterizes how well it can realize a known intent through its frozen NATIVE

interface. The REVEAL route used in G_s is instead the common intent-revealing interface used for the primary clarification comparison.

If a clarification interaction incurs cost c, we define the corresponding cost-adjusted value as

$$\mathrm{VoC}_\lambda(p, z, s) = G_s(p, z) - \lambda c,$$

where λ controls the cost assigned to an additional interaction.

3 A2I-BENCH

A2I-Bench contains 527 ambiguous text-to-image prompts designed for counterfactual evaluation of clarification. Each case pairs one prompt with two linguistically plausible and visually distinguishable interpretations, together with an intent-specific visual contract for each interpretation. The prompt itself is shared across the two user states, allowing us to vary what the user means without changing what the system initially observes.

3.1 BENCHMARK CASES

Each case consists of an ambiguous prompt p, two candidate interpretations $Z(p) = \{z_1, z_2\}$, and one visual contract for each interpretation. Both interpretations must be plausible readings of the original request before clarification and must imply observably different visual outcomes.

A2I-Bench covers four ambiguity families: *attribute scope*, *quantity*, *lexical ambiguity*, and *spatial underspecification*. Among the 527 prompts, 301 are derived from DiffusionDB (Wang et al., 2023) and 226 are controlled synthetic constructions.

All cases contain exactly two candidate interpretations. This binary design gives each prompt a controlled counterfactual pair while keeping the semantic distinction visually auditable. Figure 2 illustrates the four ambiguity families: each shared prompt admits two interpretations with distinct visual requirements. The paired images show these distinctions, while the selected positive and negative predicates make the intended outcomes explicit.

3.2 INTENT-SPECIFIC VISUAL CONTRACTS

There is generally no single reference image for an intent. We therefore represent each interpretation with a set of atomic visual constraints.

A contract contains positive predicates that should be satisfied and negative predicates that rule out the competing interpretation. Predicates cover observable objects, attributes, quantities, text, and relations. Contract satisfaction is obtained by evaluating these checks individually and aggregating them into the score $S(I, z)$ introduced in Section 2.

Negative predicates are included when satisfying a desired property alone would not rule out the competing interpretation. Across the 1,054 intent contracts, an intent contains 3.70 atomic checks on average (median 4), and 73.4% contain at least one negative requirement. This is important for ambiguity evaluation: an image should not receive full credit simply because it contains evidence for the desired interpretation while simultaneously containing the competing one.

3.3 COUNTERFACTUAL USER STATES

Each benchmark case induces two user states: one user intends z_1, and the other intends z_2. For state-blind conditions such as DIRECT and REWRITE, both users share the same generation inputs and image pools; only the intent contract used to evaluate those images changes. These user states are defined independently of the generation system. Only later, using each system's empirical DIRECT behavior, are they labeled *Direct-aligned* or *Direct-conflict*. Thus, the same fixed user intent may be Direct-aligned with one system and Direct-conflict with another.

3.4 CONSTRUCTION AND QUALITY CONTROL

A2I-Bench combines naturally occurring prompts from DiffusionDB (Wang et al., 2023) with controlled synthetic constructions. Natural prompts preserve ambiguities that occur in real text-to-image

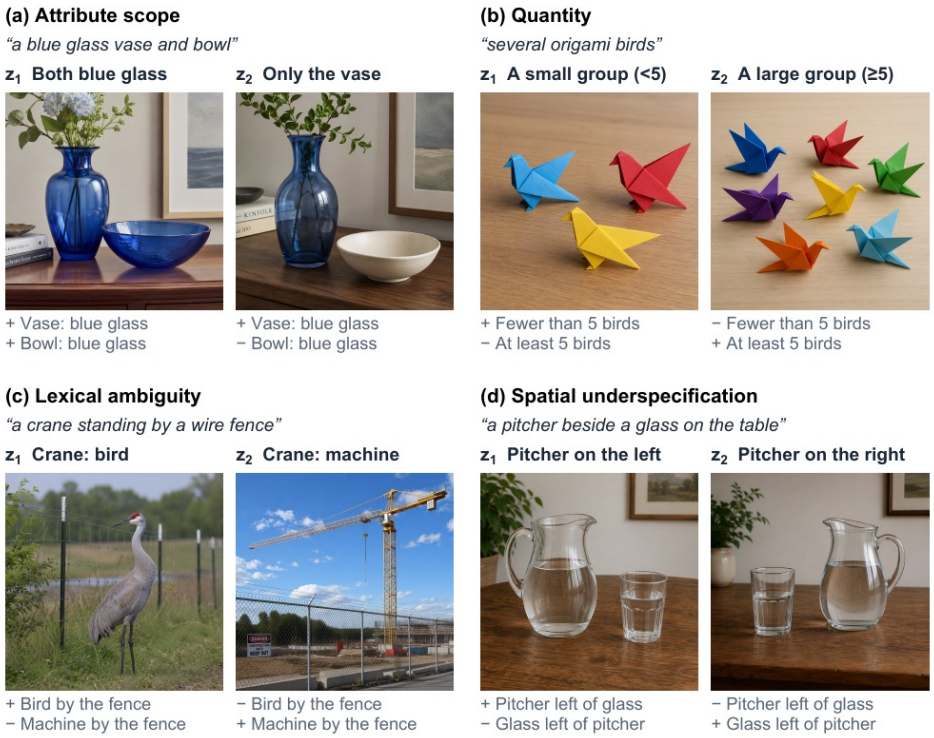


Figure 2: Selected A2I-Bench cases spanning four ambiguity families. Each original prompt is paired with two fixed intents and selected atomic requirements (+: the predicate must hold; −: it must not hold). Other contract checks are omitted for readability. Images are existing GPT-Image-2 outputs under Reveal, selected for visual clarity and composition.

requests, while synthetic cases provide targeted coverage of ambiguity structures that are otherwise sparse. Candidate cases are retained only when both interpretations are linguistically plausible and imply distinguishable visual outcomes.

Related cases are tracked by construction group. The final 527 cases span 396 such groups, which are used for statistical aggregation and uncertainty estimation. All 527 cases have complete generation support for all four evaluated systems, yielding $527 \times 4 = 2{,}108$ prompt-system units and $4{,}216$ prompt-system-user instances. Appendix B provides further construction and quality-control details. We further validate that the two benchmark interpretations correspond to natural competing user intents through a human study in Appendix B.1.

4 Evaluation Protocol

Our primary analysis asks whether revealing a user's intended interpretation is more valuable when interpretation conflicts with the generator's independently estimated Direct preference. The main challenge is that the default itself must be estimated from samples. We therefore separate the samples used to infer a system's default behavior from the outcomes used to measure clarification gain.

4.1 Generation systems

We evaluate four text-to-image systems spanning API and open-weight models: **GPT-Image-1.5** (OpenAI, 2025), **GPT-Image-2** (OpenAI, 2026), **FLUX.1-dev** (BlackForestLabs, 2024), and **Qwen-Image** (Wu et al., 2025). We treat each model together with its conditioning interface as a generation stack.

Before the scale evaluation, we freeze one resolved-intent interface for each stack. GPT-Image-1.5 and GPT-Image-2 use contrastive conditioning, FLUX.1-dev uses positive-only conditioning, and Qwen-Image uses its native positive/negative interface. Interface selection and generation settings are detailed in Appendix C.

4.2 Controlled generation conditions

To separate default interpretation, clarification gain, and resolved-intent execution, we evaluate four controlled generation conditions.

Direct. The generator receives the original ambiguous prompt. We draw 16 independent samples and use them to estimate which interpretation the system favors in the absence of additional information.

Rewrite. A language model receives only the original prompt and produces a single self-contained image prompt without asking a question. It must commit to its best reading of the request and cannot observe the candidate intents or visual contracts. We generate four images from this frozen rewrite.

Reveal. The intended interpretation is supplied explicitly, and four images are generated from a common intent-revealing representation.

Native. The same resolved intent is expressed through its frozen generator-specific interface, again with four images. Native measures how well each stack realizes an intent through its preferred interface.

The primary clarification comparison is therefore Reveal versus Rewrite. Both conditions use four images per intent or prompt, so their difference does not depend on a larger generation budget. Direct is used to estimate the generator's default interpretation, while Native measures resolved-intent execution capability.

4.3 Cross-fitted direct defaults

A finite set of Direct samples gives a noisy estimate of the system's Direct preference. Using the same estimate indiscriminately can exaggerate the distinction between Direct-aligned and Direct-conflict states. We therefore determine aligned and conflict states by cross-fitting.

For each prompt-system pair, we partition the 16 Direct samples into two fixed folds of eight images, A and B. From fold A, we estimate

$$z^{\star}_{s,A}(p) = \arg\max_{z \in Z(p)} \frac{1}{|A|} \sum_{I \in A} S(I, z), \tag{10}$$

and analogously obtain $z^{\star}_{s,B}(p)$ from fold B.

Each fold independently labels the two fixed user intents as *aligned* or *conflict*. We compute the corresponding clarification gains and average the two fold directions.

These labels are defined independently of the Rewrite outcomes used in the primary clarification comparison. Because Rewrite can itself disambiguate the prompt during reformulation, its realized interpretation can differ from the Direct-defined preference; we examine this distinction in Appendix E.5.

4.4 Matched outcome measurement

Direct and Rewrite are generated once per prompt-system pair and are state-blind: counterfactual users differ only in which intent contract scores these shared outputs. Reveal conditions on the corresponding intended interpretation instead. For intent z, the primary per-instance gain is

$$G_s(p, z) = \overline{S}_{\mathrm{Reveal}}(p, z, s) - \overline{S}_{\mathrm{Rewrite}}(p, z, s). \tag{11}$$

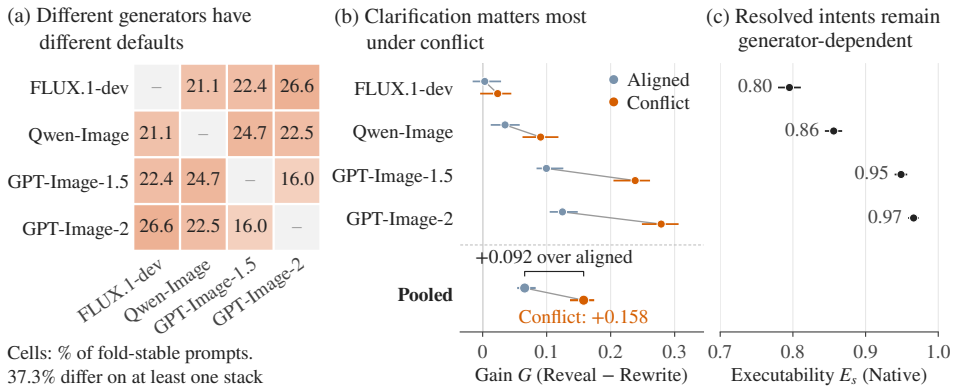


Figure 3: Why clarification value is generator-dependent. **(a)** Pairwise disagreement in default interpretation among prompts whose Direct preference is fold-stable for both systems; among prompts stable for all four systems, at least one default differs on 37.3% (136/365). **(b)** Reveal improves contract satisfaction more in the conflict state for every stack; pooled conflict gain is 0.158, which is 0.092 larger than the aligned gain. Error bars are descriptive two-sided 95% bootstrap intervals; the corresponding one-sided lower bounds for the pooled conflict gain and conflict–aligned difference are 0.140 and 0.067. **(c)** Resolved-intent executability remains stack-dependent under each system's frozen Native interface.

We then relabel the two fixed intents using the cross-fitted generator default and estimate the two quantities introduced in Section 2: the mean gain in the conflict state, ν, and the conflict–aligned difference, Δ.

4.5 Aggregation and uncertainty

We evaluate all 527 A2I-Bench prompts on each of the four generation systems, yielding 2,108 prompt–system units across 396 construction groups. Because related prompts can arise from the same underlying construction, we first average cases within each construction group and assign equal weight to each group. We estimate uncertainty with a two-level bootstrap: construction groups are resampled at the outer level, and the image pools entering each sampled prompt-stack unit are resampled at the inner level. Appendix D gives the full procedure.

4.6 Contract evaluation

Each generated image is evaluated against the atomic predicates in its intent-specific visual contract using a frozen multimodal judging pipeline. The resulting predicate judgments are aggregated into the contract score $S(I, z)$.

We separately evaluate the reliability of these atomic judgments using human annotations. Across 285 items labeled by both annotators, human agreement is 90.5%; the automatic evaluator has error rates of 11.0% and 6.6% against the two annotators, respectively. Judge implementation and measurement validation are detailed in Appendix C.7 and F.

5 Results

The experiments reveal a simple pattern: clarification is most useful when the user's intended interpretation is not the one the generation system would otherwise choose. The size of this effect, however, varies substantially across generators.

5.1 Different generators prefer different interpretations

The notion of a "default interpretation" is not shared across generation systems (Figure 3(a)). Among the 365 A2I-Bench cases for which the preferred interpretation is stable across both Direct folds for all four systems, at least two systems disagree on the preferred interpretation for 136

prompts (37.3%). Pairwise disagreement on this fold-stable subset ranges from 16.0% between GPT-Image-1.5 and GPT-Image-2 to 26.6% between FLUX.1-dev and GPT-Image-2. Thus, an ambiguous prompt does not have a generator-independent default reading. Whether a particular user intent is already aligned with the system depends on which system is generating the image.

5.2 Clarification gain varies with no-question alignment

We evaluate the central matched comparison (Figure 3(b)). For each prompt and system, the same two fixed user intents are labeled as Direct-aligned or Direct-conflict using the cross-fitted Direct preference. We then measure the gain from Reveal over the frozen Rewrite baseline.

Across all four systems, revealing the intended interpretation improves contract satisfaction by $\nu = \mathbf{0.158}$ in the Direct-conflict state. The corresponding gain in the Direct-aligned state is 0.066, yielding a conflict-aligned difference of $\Delta = \mathbf{0.092}$. As a sensitivity analysis, we instead define alignment using held-out Rewrite behavior itself. The contrast remains positive and becomes larger: the Rewrite-defined conflict gain is 0.213, the aligned gain is 0.010, and $\Delta = 0.203$; see Appendix E.5. Together, these analyses show that the conflict–aligned contrast is most directly associated with whether the no-question route already realizes the user's intent. The independently estimated Direct preference provides a useful but imperfect proxy for this alignment.

The effect is not uniform across systems. GPT-Image-1.5 and GPT-Image-2 show large gains in the conflict state (0.238 and 0.279), whereas the corresponding gains are smaller for Qwen-Image (0.091) and FLUX.1-dev (0.023). The conflict-aligned contrast follows the same pattern. Therefore, the value of clarification depends on the system that receives the clarified intent.

Together, these analyses show that the conflict–aligned contrast primarily tracks whether the no-question route already realizes the user's intent. A headroom analysis in Appendix E.8 further shows that this contrast largely disappears after conditioning on held-out Rewrite satisfaction. We therefore use aligned/conflict states as retrospective diagnostics of where clarification gain arises, rather than as ask-time decision variables.

5.3 Knowing the intent does not remove execution differences

Why does clarification help some systems more than others? Figure 3(c) shows one reason: resolving the ambiguity and rendering the resolved interpretation are separate problems.

Using each system's frozen native resolved-intent interface, mean contract satisfaction is 0.795 for FLUX.1-dev, 0.856 for Qwen-Image, 0.949 for GPT-Image-1.5, and 0.966 for GPT-Image-2. These differences remain even though the intended interpretation is already known. The gap is particularly visible on compositional constraints. For quantity cases, resolved-intent satisfaction ranges from 0.745 for FLUX.1-dev to 0.922 for GPT-Image-2; for spatial underspecification, it ranges from 0.751 to 0.989. Clarification can remove uncertainty about what should be drawn, but it cannot remove limitations in drawing it.

5.4 The main effect is stable under stricter analyses

The effect is stronger on fold-stable units: $\nu = 0.166$ and $\Delta = 0.099$, with positive lower confidence bounds (0.145 and 0.069). Additional analyses in Appendix E examine preference-label stability, stronger no-question baselines, and interaction cost.

6 Conclusion

We introduce A2I-Bench, a benchmark for measuring the value of clarification in interactive text-to-image generation. A2I-Bench shows that ambiguity alone does not determine whether clarification will improve the generated image, and generators also differ in their default interpretations and resolved-intent execution capabilities. These results suggest a more useful target for interactive generation than ambiguity detection alone. The relevant quantity is whether revealing the missing information changes what the no-question system would otherwise produce for the user. A2I-Bench provides a controlled setting for measuring that quantity and for studying clarification policies that condition on both user uncertainty and generator behavior.

AI USE STATEMENT

In this work, we used generative AI tools for code development and debugging, feedback on experimental design, and manuscript refinement. All AI-assisted code was executed and verified by the authors, all reported numerical results were checked against experiment outputs, and all technical claims and citations were manually reviewed. The authors take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the benchmark construction, analyses in the main paper and appendix. Code and the benchmark will be released to facilitate reproduction of the reported results.

REFERENCES

BlackForestLabs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Jasin Cekinmez, Addison J Wu, Raja Marjieh, and Thomas L Griffiths. Where did the ambiguity go? examining how multimodal models interpret polysemous words. *arXiv preprint arXiv:2608.00410*, 2026.

Chieh-Yun Chen, Min Shi, Gong Zhang, and Humphrey Shi. T2i-copilot: A training-free multi-agent text-to-image system for enhanced prompt interpretation and interactive generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19396–19405. IEEE, 2025.

Zijie Chen, Lichao Zhang, Fangsheng Weng, Lili Pan, and Zhenzhong Lan. Tailored visions: Enhancing text-to-image generation with personalized prompt rewriting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7727–7736. IEEE, 2024.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for step-by-step text-to-image generation and evaluation. *Advances in Neural Information Processing Systems*, 36:6048–6069, 2023.

Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International conference on learning representations*, volume 2024, pp. 15625–15645, 2024.

Yijiang River Dong, Tiancheng Hu, Zheng Hui, Caiqi Zhang, Ivan Vulić, Andreea Bobu, and Nigel Collier. Value of information: A framework for human–agent communication. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42879–42896, 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

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. *arXiv preprint arXiv:2412.06771*, 2024.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939, 2023.

Yangfan He, Jianhui Wang, Yijin Wang, Yan Zhong, Xinyuan Song, Junjiang Lin, Xinhang Yuan, Jingqun Tang, Yi Xin, Hao Zhang, et al. Enhancing intent understanding for ambiguous prompt: A human-machine co-adaption strategy. *arXiv preprint arXiv:2501.15167*, 2025.

Nailei Hei, Qianyu Guo, Zihao Wang, Yan Wang, Haofen Wang, and Wenqiang Zhang. A user-friendly framework for generating model-preferred prompts in text-to-image synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2139–2147, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

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 *2023 ieee/cvf international conference on computer vision (iccv)*, pp. 20349–20360. IEEE, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 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. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011, 2023.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 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 *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Ninareh Mehrabi, Palash Goyal, Apurv Verma, Jwala Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel, Aram Galstyan, and Rahul Gupta. Resolving ambiguities in text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14367–14388, 2023.

Ofir Nabati, Guy Tennenholtz, ChihWei Hsu, Moonkyung Ryu, Deepak Ramachandran, Yinlam Chow, Xiang Li, and Craig Boutilier. Preference adaptive and sequential text-to-image generation. *arXiv preprint arXiv:2412.10419*, 2024.

OpenAI. The new chatgpt images is here. `https://openai.com/index/new-chatgpt-images-is-here/`, December 2025.

OpenAI. Introducing chatgpt images 2.0. `https://openai.com/index/introducing-chatgpt-images-2-0/`, April 2026.

Cheng Qian, Bingxiang He, Zhong Zhuang, Jia Deng, Yujia Qin, Xin Cong, Zhong Zhang, Jie Zhou, Yankai Lin, Zhiyuan Liu, et al. Tell me more! towards implicit user intention understanding of language model driven agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1088–1113, 2024.

Sudha Rao and Hal Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2737–2746, 2018.

Michael Saxon, Fatima Jahara, Mahsa Khoshnoodi, Yujie Lu, Aditya Sharma, and William Yang Wang. Who evaluates the evaluations? objectively scoring text-to-image prompt coherence metrics with t2iscorescore (ts2). *Advances in Neural Information Processing Systems*, 37:85630–85657, 2024.

Jianwen Sun, Yukang Feng, Yifan Chang, Chuanhao Li, Zizhen Li, Jiaxin Ai, Fanrui Zhang, Yu Dai, and Kaipeng Zhang. Dialogue as discovery: Navigating human intent through principled inquiry. *arXiv preprint arXiv:2510.27410*, 2025.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, Aoxue Li, and Xihui Liu. T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8919–8944, 2026.

Manan Suri, Puneet Mathur, Nedim Lipka, Franck Dernoncourt, Ryan A Rossi, and Dinesh Manocha. Structured uncertainty guided clarification for llm agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 40811–40838, 2026.

Linqing Wang, Zhiyong Xu, Ximing Xing, Yiji Cheng, Zhiyuan Zhao, Donghao Li, Tiankai Hang, Zhenxi Li, Jiale Tao, Qixun Wang, et al. Promptenhancer: Taming your rewriter for text-to-image generation via fine-grained reward. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14895–14904, 2026a.

Wenxuan Wang, Shi Juluan, Zixuan Ling, Yuk-Kit Chan, Chaozheng Wang, Cheryl Lee, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. Learning to ask: When llm agents meet unclear instruction. In *Proceedings of the 2025 conference on empirical methods in natural language processing*, pp. 21784–21795, 2025.

Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Man Zhang, and Xiangxiang Chu. Everything in its place: Benchmarking spatial intelligence of text-to-image models. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=ddFN3lWpIr.

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 893–911, 2023.

Xinyi Wen, Lena Hegemann, Xiaofu Jin, Shuai Ma, and Antti Oulasvirta. Adaptive prompt elicitation for text-to-image generation. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*, pp. 730–754, 2026.

Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Pinelopi Papalampidi, Ira Ktena, Christopher Knutsen, et al. Revisiting text-to-image evaluation with gecko: on metrics, prompts, and human rating. In *International Conference on Learning Representations*, volume 2025, pp. 272–287, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Mingrui Wu, Lu Wang, Pu Zhao, Fangkai Yang, Jianjin Zhang, Jianfeng Liu, Yuefeng Zhan, Weihao Han, Hao Sun, Jiayi Ji, et al. Reprompt: Reasoning-augmented reprompting for text-to-image generation via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 14030–14057, 2026.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

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. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Shih-Ying Yeh, Yi Li, SangHyun Park, Giyeong Oh, Xuehai Wang, Min Song, Youngjae Yu, and Shang-Hong Lai. Tipo: Text to image with text pre-sampling for prompt optimization. In *International Conference on Learning Representations*, volume 2026, pp. 131998–132048, 2026.

Jingtao Zhan, Qingyao Ai, Yiqun Liu, Jia Chen, and Shaoping Ma. Capability-aware prompt reformulation learning for text-to-image generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2145–2155, 2024.

Michael Zhang, W Bradley Knox, and Eunsol Choi. Modeling future conversation turns to teach llms to ask clarifying questions. In *International Conference on Learning Representations*, volume 2025, pp. 60722–60742, 2025.

Michael JQ Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with lms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5541–5558, 2025.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. Clamber: A benchmark of identifying and clarifying ambiguous information needs in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10746–10766, 2024.

Zekai Zhang, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiaoyue Chen, et al. Qwen-image-agent: Bridging the context gap in real-world image generation. *arXiv preprint arXiv:2606.26907*, 2026.

Jiale Zhao, Ke Fang, and Lu Cheng. When and what to ask: Askbench and rubric-guided rlvr for llm clarification. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17120–17140, 2026.

A Related Work

Clarification and value of information. Clarification can be viewed as a decision about whether acquiring additional user information is worth the interaction cost. Early work ranks clarification questions using expected value of perfect information (Rao & Daumé III, 2018), while more recent benchmarks and agentic methods study ambiguity detection, intent elicitation, and when or what to ask (Zhang et al., 2024; Qian et al., 2024; Zhang & Choi, 2025; Zhang et al., 2025; Sun et al., 2025; Wang et al., 2025; Zhao et al., 2026). Recent formulations make this utility perspective explicit, selecting questions according to expected information value or trading expected downstream utility against the cost of interrupting the user (Suri et al., 2026; Dong et al., 2026). These works establish that uncertainty alone does not justify interaction. A2I-Bench studies an additional source of variation in generative vision: the user's intended interpretation may already coincide with what a particular generator would produce without clarification, and even a revealed intent may remain difficult for that generator to realize. We therefore measure clarification value conditional on the downstream generator rather than treating it as a property of the ambiguous request alone.

Interactive and preference-aware text-to-image generation. Ambiguous and underspecified prompts have motivated a growing line of interactive text-to-image systems. Mehrabi et al. introduce TAB and TIED for resolving multiple plausible interpretations through clarification or explicit alternatives (Mehrabi et al., 2023), while Proactive T2I Agents (Hahn et al., 2024) actively elicit missing information over multiple turns. More recent systems broaden the interaction loop through iterative refinement, planning, retrieval, memory, and feedback (Chen et al., 2025; Zhang et al., 2026), while related work adapts generation sequentially to user preferences or actively elicits latent visual requirements (Nabati et al., 2024; Wen et al., 2026; He et al., 2025). These methods primarily study how user information should be acquired or incorporated into generation. A2I-Bench asks a complementary question: for a fixed candidate intent, how much does revealing that intent actually change the visual outcome for a particular generator? Its matched counterfactual design isolates this downstream value under a fixed generation budget.

Prompt reformulation without user interaction. Clarification is not the only way to improve an underspecified request: a system can reformulate the prompt without obtaining new information from the user. Promptist (Hao et al., 2023) adapts user inputs toward prompts preferred by a downstream text-to-image model, and subsequent work develops capability-aware, user-friendly, and personalized reformulation (Zhan et al., 2024; Hei et al., 2024; Chen et al., 2024). More recent approaches optimize prompt rewriting through text presampling, reasoning, reinforcement learning, or fine-grained visual rewards (Yeh et al., 2026; Wu et al., 2026; Wang et al., 2026a). This literature motivates a strong no-question comparator: better prompt formulation can improve generation without revealing the user's latent intent. Accordingly, A2I-Bench's Rewrite condition may improve the original wording but does not observe which candidate interpretation the user intends, separating the value of additional user information from the value of prompt reformulation itself.

Generator behavior and fine-grained visual evaluation. Text-to-image generators differ not only in overall image quality but also in their semantic and compositional behavior. Recent work shows that models can favor different senses of ambiguous language (Cekinmez et al., 2026), while broader evaluations document substantial capability variation across generation systems (Lee et al., 2023). Global alignment and preference metrics such as CLIPScore (Hessel et al., 2021), ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), and HPSv2 (Wu et al., 2023) provide useful aggregate signals, but do not directly identify which fine-grained visual requirements are satisfied. Structured evaluators and benchmarks instead decompose text-image alignment into objects, attributes, relations, questions, or other localized semantic checks (Hu et al., 2023; Ghosh et al., 2023; Huang et al., 2023; Cho et al., 2023; 2024; Lin et al., 2024; Li et al., 2024). Recent work further examines the reliability and saturation of automatic evaluation and extends testing to harder spatial and reasoning-sensitive settings (Saxon et al., 2024; Wiles et al., 2025; Kamath et al., 2025; Wang et al., 2026b; Sun et al., 2026). A2I-Bench adopts the same principle of fine-grained semantic evaluation but changes the estimand: the same generated output can be scored against competing, intent-specific visual contracts, enabling separate measurement of what a generator tends to produce before resolution and what it can realize once the intended interpretation is known.

B BENCHMARK CONSTRUCTION

Figure 4 summarizes how we construct A2I-Bench. We begin with 2,880 seed prompts from DiffusionDB (Wang et al., 2023) and controlled synthetic constructions. Language-model proposers expand these seeds into 7,006 candidate ambiguity cases, which are progressively curated for ambiguity validity, redundancy, and evaluability. A pre-specified data partition determines the evaluation pool, after which cases without complete support across the four generation systems are removed. The resulting A2I-Bench contains 527 prompts.

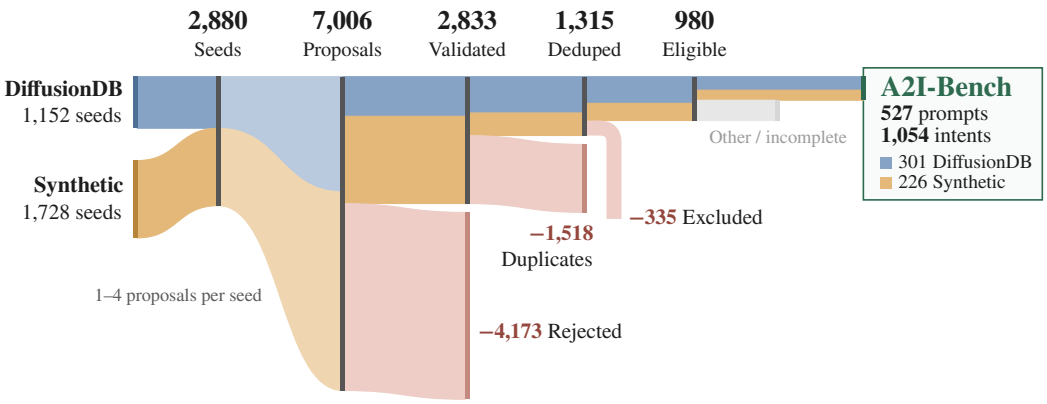


Figure 4: Construction of A2I-Bench. Natural and synthetic seed prompts are expanded into candidate ambiguity cases and progressively curated for ambiguity validity, redundancy, and evaluability. A pre-specified construction-group partition defines the evaluation pool. The final benchmark contains 527 prompts with two candidate intents each.

Candidate generation. We collect 1,152 natural seed prompts from DiffusionDB (Wang et al., 2023) and construct 1,728 synthetic seeds to improve coverage of controlled ambiguity patterns. Multiple language-model proposers convert these 2,880 seeds into candidate cases containing two plausible visual interpretations and their associated visual contracts, yielding 7,006 candidate cases.

Curation. Candidates are screened for whether the proposed ambiguity is genuine, visually distinguishable, and suitable for contract-based evaluation. We additionally remove semantic near-duplicates and cases that overlap development data or cannot be evaluated consistently across the generation systems. After these checks, 980 eligible cases remain.

Final benchmark. We retain cases with complete, scorable support on all four generation systems, resulting in the final 527-case A2I-Bench. The benchmark contains 301 DiffusionDB-derived and 226 synthetic prompts, spans 396 construction groups, and defines 1,054 intent-specific visual contracts.

B.1 HUMAN INTENT VALIDITY

We conduct a human study to verify that the two interpretations in each A2I-Bench case correspond to natural competing latent user intents rather than merely constructed alternatives. We use two disjoint annotator pools, with five annotators per prompt in each pool.

Blind intent elicitation. Annotators in the first pool see only the original prompt. They first indicate whether the prompt admits more than one reasonable visual interpretation and freely describe the image they would expect to be generated. After this response is locked, the two benchmark interpretations are shown in randomized order, and annotators indicate whether their expectation is closer to z_1, z_2, neither (*other*), or whether they have no preference. The free-form descriptions are also independently mapped to z_1, z_2, *other*, or *uncommitted* by a second coder.

Subset	n	Ambig. (%)	Both plausible (%)	Blind both (%)	Minority $P(z \mid p)$	Contract valid (%)
All	481	91.1	87.8	65.2	0.34	95.4
DiffusionDB	277	92.4	86.8	64.7	0.33	95.5
Synthetic	204	88.6	89.2	67.0	0.34	95.3
Attribute scope	254	91.0	86.0	61.7	0.31	96.1
Quantity	112	92.2	87.7	68.9	0.37	94.2
Lexical	23	89.2	76.7	40.0	0.26	89.1
Spatial	92	90.2	95.3	76.1	0.39	96.7

Table 1: Human validation of the two candidate intents in A2I-Bench. **Ambig.** is the fraction of prompts judged by a majority to admit multiple reasonable visual interpretations. **Both plausible** is the fraction for which both benchmark interpretations receive a plausibility rating of at least 4 from at least three of five annotators. **Blind both** is the fraction for which both benchmark interpretations appear among blind intent responses. **Minority** $P(z \mid p)$ is the mean empirical probability assigned to the less common benchmark interpretation among committed blind responses. **Contract valid** is the fraction of intent contracts judged free of major issues.

These blind choices provide an empirical estimate of the latent intent distribution

$$\hat{P}(z \mid p) = \frac{N_p(z)}{N_p(z_1) + N_p(z_2)}, \qquad z \in \{z_1, z_2\}, \tag{12}$$

where *other* and *no preference/uncommitted* responses are retained and reported separately rather than forced into one of the two benchmark intents.

Intent plausibility. A separate annotator pool sees the original prompt together with the two candidate interpretations and independently rates whether each interpretation is a plausible reading of the request on a five-point scale. We consider an interpretation plausible when at least three of five annotators assign a score of at least 4. A case is considered to contain two plausible competing intents when both interpretations satisfy this criterion. Annotators additionally judge whether the two interpretations are visually distinguishable and whether the associated visual contracts faithfully represent their corresponding intents.

Results. Across the 481 non-foil benchmark cases, a majority of annotators judged 91.1% of prompts to admit multiple reasonable visual interpretations. Both benchmark intents were independently judged plausible for 87.8% of cases. Importantly, the two interpretations were not only accepted after being shown to annotators: both appeared in blind intent elicitation for 65.2% of prompts. Among annotators who committed to one of the two benchmark readings, the mean minority-intent probability was 0.34, indicating substantial mass on the less common interpretation rather than a single effectively deterministic reading.

Table 1 summarizes the results by benchmark source and ambiguity family.

The individual intent plausibility scores are also high and balanced: the mean ratings are 4.28 for z_1 and 4.35 for z_2. The corresponding majority-plausibility rates are 94.3% and 93.4%, respectively. Thus, the aggregate "both plausible" result is not driven by one consistently dominant interpretation.

Controls and annotation quality. We include unambiguous prompts and deliberately implausible interpretations as controls. Only 2.1% of 97 unambiguous control prompts are judged ambiguous by a majority of annotators, while only 1.3% of deliberately implausible foil interpretations are accepted as plausible. Independent coding of the blind free-form descriptions agrees in 91.6% of cases. Table 2 summarizes these checks.

Overall, these results support the intended interpretation of A2I-Bench as a collection of prompts with two natural, competing latent user intents. The blind responses further provide a prompt-specific empirical prior $\hat{P}(z \mid p)$, while explicitly preserving users whose intended interpretation lies outside the two benchmark intents or who do not express a preference.

Measure	Value
Unambiguous controls judged ambiguous	2.1%
Implausible foils judged plausible	1.3%
Free-description coder agreement	91.6%
Blind responses with no committed intent	15.4%
Cases with either intent judged substantially strained	10.0%

Table 2: Control and annotation-quality statistics for the human intent-validity study.

C Generation Conditions and Frozen Interfaces

C.1 Generation stacks

We evaluate four image-generation stacks:

- GPT-Image-1.5,
- GPT-Image-2,
- FLUX.1-dev,
- Qwen-Image.

A stack includes both the underlying generator and the interface used to express a resolved intent.

C.2 Frozen resolved-intent interfaces

Before scale evaluation, one resolved-intent interface was chosen for each generator and then frozen. The selected interfaces are:

Stack	Frozen interface
GPT-Image-1.5	contrastive
GPT-Image-2	contrastive
FLUX.1-dev	positive-only
Qwen-Image	native positive/negative

For GPT-Image-1.5, FLUX.1-dev, and Qwen-Image, the interface was selected on the 12-case pilot. Mean non-preferred-intent contract satisfaction in that pilot was:

Stack	Contrastive	Positive-only	Native-negative
GPT-Image-1.5	**0.966**	0.898	0.898
FLUX.1-dev	0.634	**0.760**	0.760
Qwen-Image	0.686	0.793	**0.848**

GPT-Image-2 uses the same contrastive interface as GPT-Image-1.5.

C.3 Generation conditions

For each ambiguous case, we use four conditions.

Direct. The original prompt is passed directly to the generator. The planned budget is 16 images per prompt–stack pair. These samples are used primarily to estimate the generator's preferred interpretation.

Rewrite. A frozen language-model rewriter receives only the raw user prompt and produces one self-contained image prompt. It cannot ask a question and does not receive the benchmark intents, contracts, ambiguity axis, or user state. Four images are generated from this rewrite.

Reveal. The intended interpretation is supplied through a common semantic representation and deterministically compiled into a resolved prompt. Four images are generated for each intent. This is the primary intent-revelation condition used to estimate clarification gain.

Native. The same resolved intent is expressed through the stack-specific frozen interface from Section C.2. Four images are generated per intent. Native measures the generator's resolved-intent execution capability.

The planned sampling budget is therefore

Condition	Images per relevant pool
Direct	16
Rewrite	4
Reveal	4 per intent
Native	4 per intent.

The primary treatment comparison, Reveal versus Rewrite, is budget-matched at four images on each side.

C.4 Frozen Rewrite baseline

The primary no-question comparator is generated once per prompt using Claude Sonnet 5 at temperature 0, with one candidate and no reranking. The instruction is:

> You turn a user's short image request into ONE self-contained prompt for a text-to-image model. You cannot ask questions and you will not get a reply, so commit to your single best reading of what the user most likely wants. Write the prompt as a plain description of the finished image. Do NOT enumerate alternatives, do NOT hedge, do NOT mention that the request was ambiguous, and do NOT ask anything. Output the prompt text only, with no preamble and no quotation marks.

The same frozen rewrite is used for every counterfactual user state associated with the prompt. In the benchmark evaluation, Direct and Rewrite generations are state-blind: Direct images carry no intent identifier, and a single Rewrite pool is scored separately against each possible intent.

Thus, Rewrite may improve prompt formulation while remaining blind to the user's intended interpretation.

C.5 Intent-revealed compilation

Reveal uses a deterministic compilation of the accepted intent; no language model infers the user state in the Reveal condition. The resolved representation states the desired interpretation and the visual requirements needed to distinguish it from the alternative.

Native takes the same underlying intent and expresses it using the frozen interface appropriate to the generator. The separation between Reveal and Native allows us to distinguish the value of revealing an intent from the remaining effect of generator-specific conditioning.

C.6 Sampling details

The local serving code defaults are:

- FLUX.1-dev: 1024×1024, 40 denoising steps, guidance scale 3.5, bfloat16;
- Qwen-Image: 1024×1024, 50 denoising steps, true-CFG scale 4.0, bfloat16.

GPT-Image-1.5 and GPT-Image-2 were requested at 1024×1024 using the client default `quality=low`.

C.7 CONTRACT JUDGING

We score generated images at the level of atomic visual predicates. For each image, a local Qwen3-VL-4B-Instruct judge receives the image together with all yes/no checks from both candidate intents and returns one Boolean verdict per check. The judge does not observe the original prompt, intent descriptions, expected predicate polarity, generator identity, or generation condition. Negative predicates are posed as ordinary presence questions; polarity is applied only after judging.

The frozen instruction asks the model to act as a precise visual checker, answer every labeled yes/no question about the same image, and return strict JSON Boolean outputs while judging only visible evidence.

For an intent z with atomic checks C_z, we compute

$$S(I, z) = \frac{1}{|C_z|} \sum_{c \in C_z} \mathbf{1}[\hat{y}(I, c) = \text{expect}(c)] \,.$$

All checks are weighted equally, and $S(I, z)$ is undefined if any required predicate verdict is missing.

Most predicate cells use the local verdict directly. A fixed subset of images is additionally evaluated by three remote judges under a frozen routing rule, with final verdicts obtained by calibrated voting. Across the analysis, 93.6% of predicate cells use the local judge alone, and calibrated voting changes the local verdict on 0.26% of cells.

D STATISTICAL ESTIMATION

D.1 CROSS-FITTED PREFERENCE LABELS

The aligned/conflict distinction depends on a generator's empirical Direct preference. To avoid defining this preference and evaluating it from exactly the same Direct sample set, the 16 Direct slots are divided into two fixed folds,

$$A = \{1, \ldots, 8\}, \qquad B = \{9, \ldots, 16\}.$$

For fold A, the estimated default interpretation is

$$L_A(p, s) = \arg \max_{z \in Z(p)} \frac{1}{|A|} \sum_{i \in A} S(I_i^{\text{Direct}}, z),$$

and analogously for $L_B(p, s)$.

Each fold labels the two fixed users independently. If $L_A = z_1$, then u_{z_1} is aligned and u_{z_2} is conflict for the A-defined direction; the roles reverse when $L_A = z_2$. The estimator is evaluated in both fold directions and averages the two.

Across the 527 cases, the two fold labels agree for 90.5% of prompt-stack units overall. Agreement by stack is 84.4% for FLUX.1-dev, 93.7% for GPT-Image-1.5, 93.2% for GPT-Image-2, and 90.5% for Qwen-Image.

D.2 PER-UNIT CLARIFICATION GAIN

For a fixed prompt p, stack s, and intended interpretation z, let

$$R_s(p, z) = \frac{1}{n_R} \sum_{i=1}^{n_R} S(I_i^{\text{Reveal}}, z)$$

and

$$W_s(p,z) = \frac{1}{n_W}\sum_{i=1}^{n_W} S(I_i^{\text{Rewrite}}, z),$$

with $n_R = n_W = 4$.

The primary per-intent clarification gain is

$$G_s^{(R)}(p,z) = R_s(p,z) - W_s(p,z).$$

The two fold-defined labels above map these intent-level gains into aligned and conflict states. We average the two fold directions to obtain one aligned and one conflict gain for each prompt-stack unit.

D.3 Primary estimands

The first primary quantity is the average clarification gain in the conflict state:

$$\nu_R = \mathbb{E}\left[G_s^{(R)}(p,z) \mid z \neq z_s^\star(p)\right].$$

The second is the contrast between conflict and aligned gains:

$$\Delta_R = \mathbb{E}\left[G_s^{(R)}(p,z) \mid z \neq z_s^\star(p)\right] - \mathbb{E}\left[G_s^{(R)}(p,z) \mid z = z_s^\star(p)\right].$$

Applying this estimator to the 527 A2I-Bench cases gives

$$\nu_R = 0.1578, \qquad \Delta_R = 0.0922$$

with an aligned-state gain of

$$0.1578 - 0.0922 = 0.06567.$$

D.4 Group weighting

Prompts generated from the same construction pattern are not treated as independent units with unrestricted weight. For aggregate estimates, cases are first averaged within the construction group. Construction groups then receive equal weight.

For per-stack analyses, this yields 527 cases distributed over 396 groups. Pooled four-stack summaries preserve the same group-based aggregation rather than weighting groups according to the number of prompts they happen to contain.

The grouping scheme is also used as the outer unit for resampling-based uncertainty estimates.

D.5 Bootstrap uncertainty

We estimate uncertainty with a two-level bootstrap. At the outer level, we resample construction groups with replacement. Within each sampled prompt-stack unit, the image pools entering the clarification gain are resampled before recomputing the statistic. We use $10{,}000$ bootstrap replicates.

Figure 3 reports two-sided 95% percentile bootstrap intervals. For the directional headline quantities, we additionally report one-sided 95% lower confidence bounds:

$$\nu_R = 0.1578, \qquad \text{LCB}_{95}(\nu_R) = 0.140,$$

and

$$\Delta_R = 0.0922, \qquad \mathrm{LCB}_{95}(\Delta_R) = 0.067.$$

E Additional Results and Robustness

This section expands the main result along three dimensions: variation across generation stacks, sensitivity to the estimated default interpretation, and the strength of the no-question baseline.

E.1 Per-stack clarification value

The pooled result masks substantial variation across generation systems. Table 3 reports the conflict-state clarification gain and the conflict–aligned contrast separately for each stack.

	Conflict gain ν_R		Conflict–aligned gap Δ_R	
Stack	Estimate	LCB	Estimate	LCB
FLUX.1-dev	0.023	0.000	0.020	-0.018
GPT-Image-1.5	0.238	0.210	0.139	0.098
GPT-Image-2	0.279	0.254	0.154	0.118
Qwen-Image	0.091	0.067	0.056	0.020
Pooled	0.158	0.140	0.092	0.067

Table 3: Clarification value by generation stack. The primary comparator is the frozen Rewrite baseline. LCB denotes the one-sided 95% lower confidence bound.

The direction of the pooled effect is shared by all four systems, but its magnitude is strongly stack-dependent. GPT-Image-1.5 and GPT-Image-2 show the largest conflict-state gains, while FLUX.1-dev shows only a small average improvement. This difference motivates treating clarification value as a downstream system property rather than attaching one value to an ambiguous prompt in isolation.

E.2 Generators do not share a common default

We compare the empirically preferred interpretation across generation systems using Direct samples.

Table 4 restricts each pair to prompt–stack units for which both systems have stable Direct preferences across the two folds.

Pair	Disagreement	Evaluated prompts
FLUX.1-dev / GPT-Image-1.5	22.4%	420
FLUX.1-dev / GPT-Image-2	26.6%	417
FLUX.1-dev / Qwen-Image	21.1%	407
GPT-Image-1.5 / GPT-Image-2	16.0%	468
GPT-Image-1.5 / Qwen-Image	24.7%	450
GPT-Image-2 / Qwen-Image	22.5%	445

Table 4: Cross-generator disagreement in default interpretation. Each row includes only prompts for which both systems agree with themselves across the two Direct folds.

A stronger restriction requires all four systems to have fold-stable preferences simultaneously. This leaves 365 prompts; 136 of them (37.3%) contain at least one cross-system disagreement.

If we instead require the estimated preference margin to be statistically clear for all four systems, 33 of 179 prompts (18.4%) still exhibit a disagreement.

Cross-generator disagreement persists when Direct preferences are restricted to fold-stable cases. Different systems can systematically resolve the same prompt toward different plausible visual interpretations.

E.3 RESOLVED-INTENT EXECUTABILITY

Clarification can reveal the intended interpretation without ensuring that a generator can realize it. We measure this second component with Native resolved-intent generation.

Stack	All	Attr. scope	Quantity	Spatial	Lexical
FLUX.1-dev	.795	.827	.745	.751	.911
GPT-Image-1.5	.949	.968	.889	.985	.987
GPT-Image-2	.966	.981	.922	.989	.994
Qwen-Image	.856	.864	.802	.912	.931

Table 5: Resolved-intent executability E_s. Entries are mean contract satisfaction under each system's frozen Native interface.

The overall range is substantial: 0.795 for FLUX.1-dev versus 0.966 for GPT-Image-2. Differences are especially pronounced for quantities and spatial relations. For spatial underspecification, for example, mean executability is 0.751 for FLUX.1-dev and 0.989 for GPT-Image-2.

This measures a different stage of the pipeline: performance after the intended interpretation has been supplied. A generator can favor the competing interpretation under Direct generation but execute the intended one reliably once it is revealed; conversely, an accurate clarification can have limited value when the resolved visual constraint remains difficult to render.

E.4 PREFERENCE-LABEL STABILITY

The primary aligned/conflict labels are obtained by cross-fitting the Direct generations. We evaluate two alternatives.

First, restricting to prompt–stack units for which both Direct folds choose the same interpretation leaves 1,907 units. On this subset,

$$\nu_R = 0.166 \qquad \text{and} \qquad \Delta_R = 0.099,$$

with lower confidence bounds 0.145 and 0.069, respectively.

Second, estimating the preferred interpretation in-sample from all 16 Direct images produces

$$\nu_R^{\text{in}} = 0.1611, \qquad G_{\text{aligned}}^{\text{in}} = 0.0624,$$

compared with

$$\nu_R^{\text{crossfit}} = 0.1578, \qquad G_{\text{aligned}}^{\text{crossfit}} = 0.0657.$$

The separation between conflict and aligned states is 0.0987 under the in-sample labels and 0.0922 under cross-fitting.

The effect therefore persists when preference estimation is restricted to the most stable examples and when the preference label is estimated from a larger Direct pool.

E.5 DIRECT- VERSUS REWRITE-DEFINED ALIGNMENT

Our primary analysis defines alignment using DIRECT generations. This choice provides an independent estimate of how the generation stack interprets the original ambiguous prompt before any prompt reformulation. The primary clarification comparison, however, is REVEAL− REWRITE. Because the rewriter may itself resolve part of the ambiguity, the interpretation realized by REWRITE need not always coincide with the DIRECT-defined preference. We therefore examine their agreement and repeat the analysis using REWRITE-defined alignment.

Analysis	Conflict gain ν	Aligned gain	Δ
Direct-defined (primary)	0.158	0.066	0.092
Rewrite-defined, cross-fitted	0.213	0.010	0.203
Direct = Rewrite	0.265	-0.007	0.272
Direct $\neq$ Rewrite	-0.059	0.269	-0.328

Table 6: Relation between DIRECT- and REWRITE-defined alignment. The first two rows compare the primary DIRECT-defined analysis with an alternative cross-fitted labeling based on held-out REWRITE behavior. The last two rows decompose units according to whether the two routes prefer the same interpretation.

Agreement between Direct and Rewrite. For each prompt–stack unit, we define the REWRITE preference analogously to the DIRECT preference, by comparing mean contract satisfaction under the two candidate interpretations. In the full pool, DIRECT and REWRITE select the same interpretation for 80.3% of the 1,623 units for which neither preference is tied. Ties are more frequent for REWRITE: 399 of 2,108 units are tied, compared with 220 for DIRECT, reflecting the smaller four-image REWRITE pool.

To avoid using the same REWRITE images both to define the preference label and to measure the no-question outcome, we additionally construct a cross-fitted analysis. The four REWRITE images are partitioned into two fixed folds of two. In each direction, one fold determines the preferred interpretation and the other supplies the no-question contract score; the two directions are then averaged. Among untied unit–design pairs, the cross-fitted DIRECT and REWRITE labels agree in 81.3% of cases.

Rewrite-defined alignment. We next replace the DIRECT-defined aligned/conflict labels with the cross-fitted REWRITE-defined labels. The resulting conflict–aligned contrast remains positive and becomes larger. As shown in Table 6, the REWRITE-defined conflict gain is 0.213, compared with an aligned gain of 0.010, giving

$$\Delta_{\text{Rewrite}} = 0.203.$$

For comparison, the primary DIRECT-defined analysis gives a conflict gain of 0.158, an aligned gain of 0.066, and

$$\Delta_{\text{Direct}} = 0.092.$$

Agreement and disagreement subsets. The decomposition in Table 6 clarifies the relation between the two definitions. When DIRECT and REWRITE prefer the same interpretation, the conflict–aligned difference is large and positive ($\Delta = 0.272$). When they prefer different interpretations, the sign reverses ($\Delta = -0.328$): an intent classified as conflicting with the DIRECT preference is then often the interpretation already favored by the REWRITE route, while the DIRECT-aligned intent is disfavored by REWRITE.

These results distinguish two related quantities. The DIRECT preference provides an independently measured, generator-native characterization of how the ambiguous prompt is interpreted before reformulation. The realized value of clarification, in contrast, depends most directly on what the no-question route would otherwise produce. Accordingly, DIRECT behavior serves as an informative but imperfect proxy for the alignment of the deployed no-question pipeline.

E.6 STRONGER NO-QUESTION BASELINES

Clarification competes not only with a single direct attempt but also with the option to spend additional generation compute without interacting with the user. We evaluate this by replacing the no-question outcome with increasingly large best-of-k Direct sets.

As k increases, the conflict-state gain decreases from 0.200 to 0.061, while the conflict–aligned contrast remains 0.122 at $k = 8$ with an LCB of 0.106. Thus, additional sampling can substitute for part of the absolute benefit of clarification.

k	ν	LCB(ν)	Δ	LCB(Δ)
1	.200	.185	.158	.140
2	.149	.135	.159	.142
4	.102	.089	.144	.128
8	.061	.048	.122	.106

Table 7: Clarification relative to stronger Direct baselines. More no-question samples reduce the absolute advantage of clarification, while the conflict–aligned contrast remains positive across all evaluated budgets.

E.7 Cost-sensitive value

The main analysis reports clarification gain before assigning a price to interaction. For the one-question clarification studied here, setting $c = 1$. Given a per-question cost λ, the same measurements define

$$\mathrm{VoC}_\lambda = G - \lambda,$$

for a one-question clarification.

Table 8 reports this quantity over the two fixed specific-user intents, using the frozen Rewrite baseline.

Stack	$\lambda = 0$	.05	.10	.20
FLUX.1-dev	.013	-.037	-.087	-.187
GPT-Image-1.5	.169	.119	.069	-.031
GPT-Image-2	.202	.152	.102	.002
Qwen-Image	.063	.013	-.037	-.137
Pooled	.112	.062	.012	-.088

Table 8: Measured Value of Clarification under different interaction costs.

The corresponding break-even interaction cost varies substantially across generation stacks.

E.8 Retrospective Alignment and Baseline Headroom

The aligned/conflict distinction is a retrospective diagnostic: it is defined relative to the user's realized intent and therefore cannot itself be observed before clarification. We further examine whether this label explains clarification gain beyond the quality of the no-question outcome.

Without conditioning on the Rewrite baseline, the conflict coefficient equals the primary conflict–aligned contrast, 0.092. Conditioning on cross-fitted Rewrite satisfaction reduces this coefficient to -0.025 $[-0.040, -0.010]$. After additionally controlling for the Direct preference margin and Native executability, the residual coefficient is -0.002 $[-0.014, 0.009]$. A within-prompt–stack comparison between the two candidate intents similarly gives 0.003 $[-0.009, 0.015]$.

These results indicate that Direct-defined conflict primarily identifies cases in which the no-question route leaves greater headroom for improvement. Consistently, at matched Rewrite satisfaction, aligned and conflict states have nearly identical remaining clarification gain.

E.9 Prospective Clarification Policy

The analyses above explain where clarification gain arises after specifying the user's intent. We next ask whether clarification value can be predicted *before* observing that intent, using only information available at ask time.

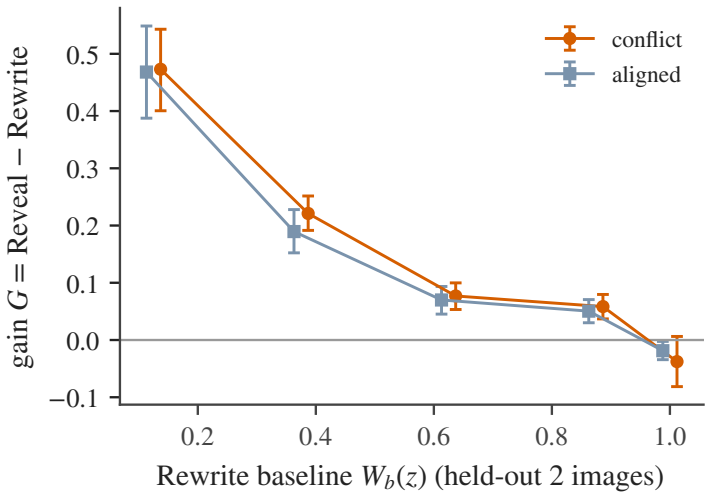


Figure 5: Clarification gain as a function of held-out Rewrite satisfaction. At matched no-question satisfaction, Direct-aligned and Direct-conflict states follow nearly identical gain curves.

Prospective target. For each prompt p and generation stack s, we treat the two benchmark intents as equally likely under the paired counterfactual design and define the expected clarification gain

$$\bar{G}(p,s) = \frac{1}{2} \sum_{z \in Z(p)} \left[R_s(p,z) - W_s^{\text{out}}(p,z) \right]. \tag{13}$$

Here $R_s(p,z)$ is the mean Reveal satisfaction for intent z, and $W_s^{\text{out}}(p,z)$ is evaluated using two held-out Rewrite samples. The remaining two Rewrite samples are available to the policy as a no-question self-check. We repeat the analysis in both two–two split directions and average the resulting estimates.

Given interaction cost λ, a learned policy asks whenever its predicted clarification gain exceeds the cost,

$$\pi_\lambda(p,s) = \mathbf{1}\left[\hat{\bar{G}}(p,s) > \lambda \right], \tag{14}$$

with policy utility

$$U_\lambda = \mathbb{E}_w \left[\pi_\lambda(p,s) \left(\bar{G}(p,s) - \lambda \right) \right], \tag{15}$$

where w denotes the same construction-group weighting used in the main evaluation.

Ask-time signals. We compare prompt-side and generator-conditioned signals. The strongest prompt/context baseline uses ambiguity family, source, LLM ambiguity votes, intent plausibility indicators, contract-complexity statistics, prompt length, stack identity, and a ridge model over text embeddings of the prompt and the two candidate intents.

Generator-conditioned signals additionally characterize the behavior of the generation stack itself. The *Direct profile* contains the mean Direct satisfaction under both candidate intents, their absolute difference, and tie information. The *self-check* evaluates the no-question Rewrite samples available to the policy against both candidate contracts and records their mean satisfaction and preference profile. The self-check therefore measures what the system would produce without clarification, while the Direct profile provides an independent characterization of its behavior under the original prompt.

Held-out evaluation. We use two evaluation protocols. First, we perform five-fold cross-validation over construction groups, so related prompts never appear in both train and test folds. Second, we perform leave-one-generator-out evaluation: the predictor is trained on three generation stacks and disjoint construction groups and evaluated on the held-out stack and held-out prompts.

Evaluation	R^2_{base}	R^2_{gen}	ΔR^2	$\Delta U_{.10}$
Group 5-fold CV	.398	.488	$+.090$ $[.045, .142]$	$+.0053$ $[.0027, .0081]$
Leave-one-generator-out	.009	.179	$+.170$ $[.116, .227]$	$+.0138$ $[.0094, .0186]$

Table 9: Prospective prediction of clarification value. Base denotes the strongest prompt/context baseline, and Gen additionally uses the Direct satisfaction profile and a no-question self-check. Utility differences are measured relative to the corresponding baseline policy at interaction cost $\lambda = 0.10$.

Reveal scores are never used as policy features. The Rewrite images used to construct W^{out} are disjoint from those used by the self-check, and Direct generations are disjoint from all outcome images. Regression hyperparameters for the text model are selected only within the training folds. Uncertainty is estimated with a paired construction-group bootstrap with 2,000 replicates.

Prospective prediction. Table 9 summarizes the main comparison. Under construction-group cross-validation, the strongest prompt/context baseline achieves a held-out R^2 of 0.398. Adding the Direct profile and self-check raises this to 0.488, an improvement of

$$\Delta R^2 = 0.090\ [0.045, 0.142].$$

At interaction cost $\lambda = 0.10$, the corresponding ask policy improves utility by

$$\Delta U_{0.10} = 0.0053\ [0.0027, 0.0081].$$

The advantage is larger under leave-one-generator-out evaluation. The prompt/context predictor achieves $R^2 = 0.009$, whereas the generator-conditioned predictor reaches $R^2 = 0.179$, giving

$$\Delta R^2 = 0.170\ [0.116, 0.227].$$

At $\lambda = 0.10$, policy utility improves by

$$\Delta U_{0.10} = 0.0138\ [0.0094, 0.0186].$$

Which generator signals matter? The improvement is not explained by the binary Direct default alone. Using the Direct profile without the self-check does not outperform the strong prompt/context baseline ($\Delta R^2 = -0.043\ [-0.084, -0.001]$). The self-check alone improves prediction by $0.050\ [0.007, 0.096]$. The two signals are complementary: adding the Direct profile on top of the self-check yields a further $0.040\ [0.021, 0.060]$ improvement in held-out R^2. Their combination is particularly important for transfer to an unseen generator, where neither signal alone provides comparable generalization.

Interpretation. These results separate two roles of generator behavior. The retrospective aligned/conflict analysis identifies where clarification gain occurred after the user intent is known, largely through the headroom left by the no-question route. The prospective analysis instead uses only information available before asking and shows that observing generator behavior improves held-out clarification decisions beyond prompt-side information alone. In particular, directly checking the no-question output is the strongest individual signal, while the Direct satisfaction profile provides complementary information about the generator and improves cross-generator transfer.

The present analysis uses the benchmark's balanced $1/2$–$1/2$ prior over its two counterfactual user states. A population-specific intent prior $q_p(z)$ can be substituted for the uniform prior in Eq. equation 13 without changing the evaluation protocol. The reported utility accounts for interaction cost but does not additionally price the computation required for Direct profiling or contract-based self-checking.

F Measurement Validation

Because the benchmark relies on fine-grained visual predicates, we evaluate the reliability of the automatic judgments with human annotations.

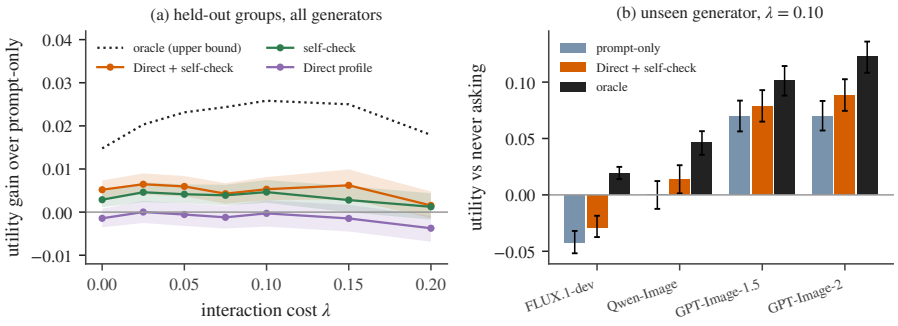


Figure 6: Prospective clarification policies. **Left:** utility improvement over the strongest prompt/context baseline under construction-group cross-validation as a function of interaction cost; the dashed curve shows the corresponding oracle gap. **Right:** leave-one-generator-out policy utility at $\lambda = 0.10$, evaluated separately on each held-out generation stack. Generator-conditioned behavioral signals improve prospective decisions under both evaluation protocols.

Human agreement. We audit 285 image-predicate pairs sampled across all four generation stacks, with the automatic verdict balanced during sampling. Two annotators independently label these items, and they agree on 258 of the 285 shared judgments (90.5%), corresponding to Cohen's $\kappa = 0.809$.

Automatic versus human judgments. Treating each annotator independently as the reference, the automatic evaluator has error rates of 11.0% and 6.6%, respectively. Table 10 further reports the corresponding sensitivity to visible positive evidence and to violations of negative predicates.

	Annotator 1	Annotator 2
Overall error	.110	.066
Positive sensitivity	.897	.940
Negative sensitivity	.667	.762

Table 10: Automatic evaluator versus human annotations at the atomic-predicate level.

Cross-provider agreement. As a complementary model-based check, we compare the local judge with an independent remote-model majority on a separately routed overlap set. They agree on $5{,}339/5{,}790 = 92.2\%$ of non-tied predicate judgments.