Seeing Isn't Believing: Uncovering Blind Spots in Evaluator Vision-Language Models

Mohammed Safi Ur Rahman Khan[♠♢], **Sanjay Suryanarayanan**[♠], **Tushar Anand**[♣], **Mitesh M. Khapra**[♠♢]

[♠]Nilekani Centre at AI4Bharat [♢] Indian Institute of Technology Madras
[♣]BITS Pilani, Hyderabad
Correspondence: {mohammed.safi, miteshk}@dsai.iitm.ac.in

https://huggingface.co/datasets/ai4bharat/Focus
https://github.com/AI4Bharat/focus

Abstract

Large Vision-Language Models (VLMs) are increasingly used to evaluate outputs of other models, for image-to-text (I2T) tasks such as visual question answering, and text-to-image (T2I) generation tasks. Despite this growing reliance, the reliability of these Evaluator VLMs remains underexplored. In this work, we systematically evaluate the reliability of Evaluator VLMs across both I2T and T2I tasks. We introduce targeted perturbations that degrade output quality along key error dimensions, including object hallucinations, spatial reasoning, factual grounding, and visual fidelity. These perturbations test whether Evaluator VLMs can reliably account for these quality degrading errors in their evaluations. Using a comprehensive benchmark of over 4000 perturbed instances spanning 40 perturbation dimensions, we evaluate 4 prominent VLMs using single-answer scoring, pairwise comparison, and reference-guided paradigms. Our findings reveal that current VLM evaluators exhibit substantial blind spots: they often fail to detect perturbed outputs - in some cases exceeding 50%, struggle particularly with fine-grained compositional and spatial errors, and are often insensitive to hallucinated content that contradicts the input image. Pairwise comparison proves more reliable, though failure rates persist. These results highlight the unreliable nature of current Evaluator VLMs and urge caution in their deployment for benchmarking and development decisions. Code and data have been made publicly available.

1 Introduction

Large Vision-Language Models (VLMs) are increasingly used to evaluate the outputs of other VLMs (Zhang et al., 2023; Yu et al., 2023; Pu et al., 2025) and image generation models (Wen et al., 2023; Zhou et al., 2025; Yang et al., 2025a) as they are scalable and cost-effective than human evaluation. Beyond benchmarking, these models are also used as reward models during training, where their feedback directly shapes model behavior (Li et al., 2025; Yasunaga et al., 2025). As a result, unreliablity of evaluator VLMs can have broad impact: they can produce misleading rankings and also reinforce undesirable behaviors during training. It is therefore important to rigorously assess their reliability.

Recent work has studied evaluator VLMs through their correlation with human judgments (Kasaei et al., 2025; Hu et al., 2025). However, establishing the dependability of evaluator VLMs requires rigorous scrutiny, to identify potential blind spots in their capabilities. For example, in image-to-text tasks such as Visual Question Answering, an evaluator must verify whether generated text is grounded in the image by detecting hallucinated objects, incorrect attributes, and fabricated facts (Jing et al., 2023; Bai et al., 2024). In text-

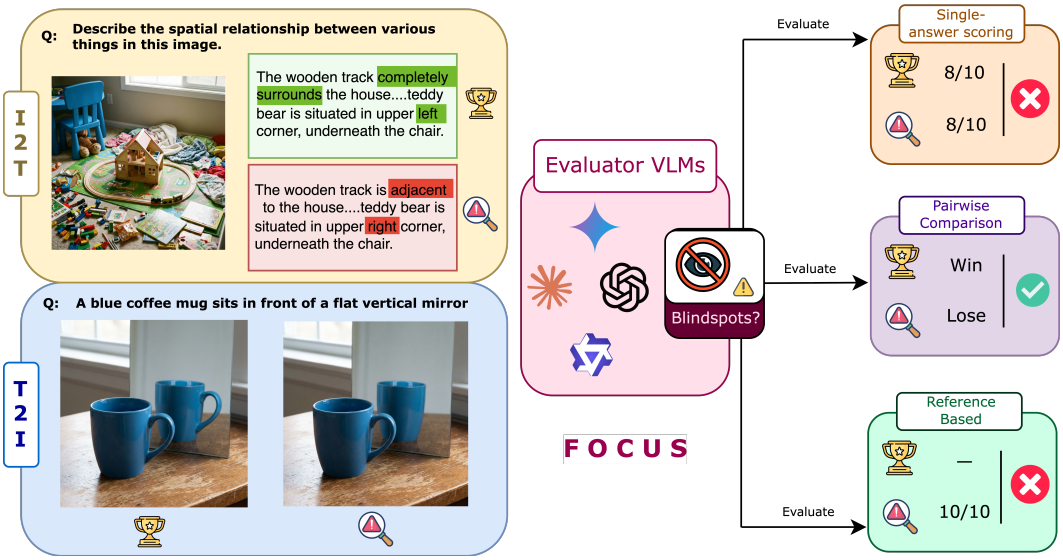


Figure 1: FOCUS is a meta-evaluation benchmark to evaluate robustness of Evaluator VLMs.

to-image tasks, it should be able to judge whether a generated image faithfully reflects the prompt, including objects, attributes, physical plausibility, and rendered text (Huang et al., 2023; Meng et al., 2024). *Can current VLMs reliably perform such fine-grained, multi-dimensional assessments, or do they exhibit systematic blind spots that render their judgments unreliable?*

In this work, we introduce FOCUS, a comprehensive meta-evaluation benchmark designed to uncover blind spots in Evaluator VLMs across both **I2T** and **T2I** tasks. Our approach is inspired by prior meta-evaluation frameworks such as Checklist (Ribeiro et al., 2020) and FBI (Doddapaneni et al., 2024) and is grounded in a simple principle: if a perturbation introduces a clear error into a model output, a reliable evaluator should detect this degradation and adjust its judgment accordingly. We design targeted perturbations spanning diverse failure modes, organized into fine-grained dimensions based on commonly reported errors in the literature (sample descriptions are shown in Table 1). Starting with 600 and 750 prompts for **I2T** and **T2I** tasks respectively, sampled from real-world benchmarks, we generate gold responses using GEMINI-3.1-PRO and GEMINI-3-PRO-IMAGE. We then introduce targeted perturbations through a rigorous human-in-the-loop process, resulting in a dataset of over 4000 instances. Each instance contains a prompt, a gold response, and a perturbed response, and is validated by expert annotators.

Using this benchmark, we evaluate four prominent VLMs under three widely adopted evaluation paradigms: single-answer scoring, pairwise comparison, and reference-guided evaluation. Within each paradigm, we explore multiple prompting strategies actually proposed in existing literature and used in practice (Ge et al., 2023; Chen et al., 2024b; Pu et al., 2025). *Our findings paint a slightly sobering picture of current VLM-based evaluation.* Evaluators frequently fail to detect quality-degrading perturbations, in some cases exceeding 50%, with performance notably worse for **T2I** than **I2T** , suggesting that current evaluators struggle with fine-grained visual understanding. Pairwise comparison emerges as the most reliable paradigm, contrasting with text-only settings where reference-based evaluation performed best (Doddapaneni et al., 2024). Failures are concentrated in categories requiring fine-grained visual grounding, compositional reasoning, and physical plausibility, and increasing reasoning budgets does not consistently help. We also notice that evaluators sometimes identify errors in justifications but fail to reflect them in final scores.

These findings have broader implications beyond evaluation. As VLMs are increasingly used as reward models during training, the blind spots we uncover suggest that reward signals from these evaluators may fail to penalize critical errors. This could inadvertently

reinforce the very behaviors they should correct. Our results highlight significant *blindspots* in current Evaluator VLMs and caution against *blind* reliance on them as standalone judges.

2 FOCUS benchmark

We introduce FOCUS, a meta-evaluation benchmark for assessing the reliability of Evaluator Vision Language Models (VLMs), or *evaluators*. Specifically, FOCUS evaluates how well evaluators assess the outputs of other VLMs and image generation models, hereby called as *evaluatees*. FOCUS consists of two splits: **I2T** (Image-to-Text) and **T2I** (Text-to-Image).

The **I2T** split covers tasks in which the *evaluatee* takes an image and a text prompt as input and produces a text response, such as Visual Question Answering (VQA) and image captioning. Each instance in this split is represented as a tuple (I, T, A_{gold}, $A_{perturb}$), where I is the input image, T is the input text prompt, A_{gold} is the correct (or *gold*) answer, and $A_{perturb}$ is a *perturbed* version of the gold answer. Similarly, the **T2I** split covers tasks in which the *evaluatee* takes a text prompt as input and produces an image, such as text-to-image generation. Each instance in this split is represented as a tuple (T, I_{gold}, $I_{perturb}$), where T is the input prompt, I_{gold} is the *gold* image, and $I_{perturb}$ is a *perturbed* version of the gold image.

In both splits, the perturbed output is created by introducing controlled errors along different perturbation dimensions (Table 1). This setup allows us to evaluate whether evaluator VLMs can reliably account for these errors during evaluation. Each instance in FOCUS is created with human oversight *throughout* the creation process, including prompt selection (§2.1), definition of perturbation dimensions (§2.2), and perturbation creation (§2.3). Detailed statistics of FOCUS are presented in Table 5 and we describe the benchmark creation process below.

2.1 Prompt selection

Each instance of FOCUS is sourced from recent popular evaluation benchmarks. For **I2T**, we manually sampled 600 instances (text prompts paired with images) from seven recent benchmarks including MMBench (Liu et al., 2023), MMDocBench (Zhu et al., 2024) and RealWorldQA[1]. We selected these benchmarks for their focus on open-ended generation tasks, an important use case for Evaluator VLMs. Similarly, for **T2I**, we manually sampled 750 instances (text prompts) from seven widely used image generation benchmarks including MJ-Bench (Chen et al., 2024b), T2I-CompBench++ (Huang et al., 2023) and T2I-ReasonBench (Sun et al., 2025a). Details of benchmarks is provided in Appendix A.3.

The gold answers (A_{gold}) and gold images (I_{gold}) were generated using GEMINI-3.1-PRO and GEMINI-3-PRO-IMAGE models, respectively and manually reviewed for quality and correctness. *Importantly, we note that perfectly accurate gold answers and images are not essential for our study, as our primary focus is on directional score changes*. Specifically, we test whether perturbed responses with clear errors receive lower scores than their corresponding original responses. Thus, it is sufficient for the gold outputs to be reasonably accurate and relevant.

2.2 Perturbation categories

Both VLMs and image generation models exhibit diverse failure modes documented in prior work (Huang et al., 2023; Meng et al., 2024; Jing et al., 2023; Bai et al., 2024). A robust evaluator should reliably detect these errors and account for them during evaluation. We therefore design perturbations grounded in these failure modes and error dimensions. Detailed descriptions of all perturbation dimensions and examples are provided in Appendix A.2. For the **I2T** split, we group perturbation dimensions into four broad categories:

Visual Grounding (VG): Perturbations that modify directly observable visual elements such as entities, attributes, spatial relations, or object presence. For example, replacing *"a dalmatian sitting on the grass"* with *"a labrador sitting on the grass"*.

[1] https://huggingface.co/datasets/xai-org/RealworldQA

Catg	Perturbation Dimension	Perturbation Description
		Image-to-Text (I2T)
VG	Entity Substitution	Swaps with a similar but incorrect entity. *Eg: The chef holds a knife → The chef holds a cleaver*
	Attribute Distortion	Changes subtle attributes like color or texture. *Eg: A red car is parked → A blue car is parked*
	Spatial Relation Swap	Alters relative positioning of objects. *Eg: A book is under a table → A book is on top of a table*
	Phantom Details Injection	Introduces non-existent objects. *Eg: A park has trees → A park has trees and a statue*
	Over Generalization	Replaces with broader hypernyms. *Eg: A woman is in a Tesla → A woman is in a vehicle*
	Important Detail Omission	Removes essential grounding elements. *Eg: a red striped hat on a table → a hat on a table*
SI	Contextual Depth Reduction	Removes implicit intent or nuance. *Eg: A contemplative man sitting → A bored man sitting*
	Cultural Misalignment	Replaces cultural markers incorrectly. *Eg: A person in a kimono → A person in a sari*
	Logical Inconsistencies	Introduces contradictions in the statement. *Eg: The open and closed bridge nearby.*
		Text-to-Image (T2I)
VF	Object Substitution	Replaces the primary object in the scene. *Eg: A cat → A dog*
	Object Addition/Omission	Alters object presence or quantity in the scene. *Eg: One chair → many chairs*
	Attribute Manipulation	Changes attributes such as color, texture, or size. *Eg: red ball → blue ball*
	Spatial Manipulation	Alters object position or relative spatial arrangement. *Eg: Cup on table → Cup under table*
	Scale Distortion	Changes object proportions or relative scale relationships. *Eg: small mouse → large mouse*
	Constraint Violation	Violates explicit prompt constraints or specified conditions. *Eg: No cars → car present*
PP	Causal Violation	Breaks expected cause-effect relationships between events. *Eg: Glass falls breaks → intact*
	Physics Manipulation	Violates basic physical laws or natural behavior. *Eg: Shadow away → towards* light
	State/Transformation Failure	Produces incorrect or incomplete transformation outcomes. *Eg: Ice melts → unchanged*
	Functional Absurdity	Depicts objects being used in illogical ways. *Eg: Knife cuts → Knife used on stone*
	Literalized Idioms	Interprets figurative expressions in a literal visual form. *Eg: Heavy rain → objects falling*

Table 1: Select perturbation categories, dimensions and examples. Original elements are shown in green and perturbed elements in red. See Table 6 for full list.

Semantic Interpretation (SI): Perturbations that degrade contextual or semantic understanding by altering cultural cues, contextual meaning, or introducing subtle inconsistencies. For example, replacing *"celebrating Diwali with diyas"* with *"celebrating Diwali with candles"*.

Visual Reasoning (VR): Perturbations that introduce logical, numerical, or causal errors. For example, replacing *"population increased by 15%"* with *"population increased by 12%"*.

Long-form Generation (LG): Perturbations that introduce inconsistencies between fluent long-form text and the underlying visual content. For example, describing *"a knight riding under the bright moon"* when the image depicts a daytime scene.

Similarly, for the **T2I** split, we define four complementary perturbation categories:

Visual Fidelity (VF): Perturbations that alter key visual elements of the generated image such as objects, attributes, or spatial layouts. For example, rendering *a red car* as *a blue car*.

Scene Coherence (SC): Perturbations that degrade scene-level consistency through stylistic mismatches, incomplete rendering, or contextual inconsistencies in the generated image. For example, generating *photorealistic humans in a flat 2D cartoon environment*.

Physical Plausibility (PP): Perturbations introducing violations of physical laws, causal logic, or common sense. For example, *shadows pointing towards the light source*.

Text Rendering (TR): Perturbations that corrupt textual or symbolic elements in a generated image. For example, *"COEFEE"* instead of *"COFFEE"* on a shop sign.

2.3 Perturbation generation

To generate perturbed responses across categories, we adopt a two-step process: automatic perturbation generation followed by thorough human verification. This approach allows us to efficiently create diverse perturbations while ensuring high quality and correctness. For **I2T** tasks, we prompt GEMINI-3.1-PRO $f(\cdot)$ with perturbation-specific instructions ($P_{perturb}$), together with the input image (I), text prompt (T), and gold answer (A_{gold}). The

model generates a perturbed answer ($A_{perturb}$) along with a short description (*descr*) of the introduced error. Formally, this process can be represented as: $f(P_{perturb}, I, T, A_{gold}) \rightarrow (A_{perturb}, descr)$. For **T2I** tasks, we first prompt GEMINI-3.1-PRO $f(\cdot)$ with perturbation-specific instructions ($P_{perturb}$), along with the text prompt (T) and the gold image (I_{gold}) to generate an edit instruction. Using the generated edit instruction, we prompt GEMINI-3-PRO-IMAGE $g(\cdot)$ to edit the gold image, and produce the perturbed image ($I_{perturb}$). Formally, this process is represented as: $g(I_{gold}, f(P_{perturb}, T, I_{gold})) \rightarrow (I_{perturb})$.

Although this automatic pipeline produces strong perturbations, all perturbed instances ($A_{perturb}$ and $I_{perturb}$) are further reviewed by human annotators, including the authors. Each perturbed response is compared with the corresponding gold response and labeled as *valid*, *invalid*, or *score-invariant*. A perturbation is marked as *valid* if it introduces a subtle but meaningful error that *should* receive a lower score than the gold response. It is marked as *invalid* if it is overly obvious or nonsensical in context. Following Doddapaneni et al. (2024), perturbations that should not receive a score penalty, such as paraphrases of the original response or image edits that do not contradict the input prompt, are categorized as *score-invariant*. To support this process, we developed a custom annotation tool. Details of the tool and annotation workflow are provided in Appendix B.

3 Experimental Setup

In this section, we first describe the prompting strategies used to benchmark Evaluator VLMs on FOCUS, and then discuss the evaluation metrics. An Evaluator VLM, $f(\cdot)$, takes as input an evaluation instruction (P_{eval}), the task input (T and/or I), and the response(s) to be evaluated (I or A), and produces a judgment along with a supporting justification. Based on existing literature, we focus on the three most commonly used evaluation paradigms: (i) Single-answer scoring, (ii) Pairwise comparison, and (iii) Reference-guided evaluation. For each paradigm, we explore commonly used prompting strategies proposed in prior works (Chen et al., 2024a; Pu et al., 2025; Lin et al., 2025; Li et al., 2024a; Yang et al., 2025b; Chen et al., 2024b; Cui et al., 2024). Our setups are directly inspired by existing literature, since our goal is to evaluate VLM Evaluators under commonly used settings rather than propose new evaluation strategies. The exact prompts and details for each evaluator are listed in the Appendix C.

Within each paradigm, we explore four prompting strategies that progressively add structure: **Vanilla [V]** (input + output only), **Rubric/Rules [R]** (adds a grading rubric or rule set), **Axes [Ax]** (evaluation along predefined axes), and **Axes+Rubric/Rules [Ax+R]** (axes with per-axis rubrics or rules). We use O_{model} to denote a single evaluatee output and O_1, O_2 for candidate pairs. Table 2 summarizes the input–output signatures for all strategies.

Single-answer scoring. In this paradigm, the evaluator scores a single response independently based on the provided input and its parametric knowledge. For **I2T** tasks, the evaluator receives the text prompt T, input image I_{in}, and generated response A_{model}; for **T2I** tasks, it receives T and the generated image I_{model}. This is the most commonly used paradigm in practice (Cui et al., 2024; Chen et al., 2024b; Pu et al., 2025).

Pairwise comparison. Here, the evaluator is tasked with selecting the better response between two candidates. For **I2T**, it receives T, I_{in}, and two candidate responses A_1 and A_2; for **T2I**, it receives T and two generated images I_1 and I_2. This paradigm is widely used in preference-based evaluation and reward modeling (Lin et al., 2025; Chen et al., 2024a).

Reference-guided scoring. In this paradigm, the evaluator scores a model output by comparing it against a reference (gold) output O_{gold}. This provides the evaluator with an explicit quality anchor, potentially simplifying judgment. However, *this approach may not be feasible for many open-ended tasks* where good references are not easily available. For **I2T**, the reference is a gold answer A_{gold}; for **T2I**, it is a gold image I_{gold}.

Paradigm	Strategy	Input to $f(\cdot)$	Output
Single-answer scoring	V	P_{eval}, T, I_{in}, O_{model}	$(score, just)$
	R	P_{eval}, R, T, I_{in}, O_{model}	$(score, just)$
	Ax	P_{eval}, $[Ax]$, T, I_{in}, O_{model}	$([score], just)$
	Ax+R	P_{eval}, $[\{Ax, R\}]$, T, I_{in}, O_{model}	$([score], just)$
Pairwise comparison	V	P_{eval}, T, I_{in}, O_1, O_2	$(verdict, exp)$
	R	P_{eval}, R, T, I_{in}, O_1, O_2	$(verdict, exp)$
	Ax	P_{eval}, $[Ax]$, T, I_{in}, O_1, O_2	$([verdict], exp)$
	Ax+R	P_{eval}, $[\{Ax, R\}]$, T, I_{in}, O_1, O_2	$([verdict], exp)$
Reference-guided scoring	Ref	P_{eval}, T, I_{in}, O_{gold}, O_{model}	$(score, just)$

Table 2: Evaluation paradigms and prompting strategies. I_{in} is included only for I2T. R denotes a rubric (single-answer) or rules (pairwise); $[Ax]$ denotes predefined evaluation axes. Brackets around output (e.g., $[score]$) indicate per-axis judgments.

3.1 Metrics

In the single-answer scoring paradigm, we measure the percentage of instances in which the score remains unchanged after perturbation (i.e., the perturbed output does not receive a strictly lower score than the gold output); unlike this paradigm, pairwise and reference-guided evaluators never see an unperturbed baseline score to compare against, so their metrics are defined below relative to what each evaluator actually observes. For evaluators that score along multiple axes, we consider only the axes relevant to the perturbation category. A perturbation is counted as undetected only if the scores on all relevant axes remain unchanged. Details about the axes are in Appendix C. Ideally, the evaluator should assign a lower score to the perturbed output. In the pairwise comparison paradigm, we present the *gold* output alongside the *perturbed* output and ask the evaluator to choose the better one. Our metric is the percentage of instances in which the evaluator fails to select the gold output alone in both order-swapped runs. For axis-based pairwise evaluators, we again restrict the analysis to the axes relevant to the perturbation category. To mitigate position bias (Zheng et al., 2023b), we run each evaluation twice, swapping the order of the gold and perturbed outputs. In reference-guided scoring, the gold output is used as the reference, and we measure the percentage of instances in which the evaluator assigns a perfect score to the perturbed output when scored against the gold reference. We use these metrics for both I2T and T2I tasks, since the evaluation paradigms are structurally identical. For score-invariant perturbations, the desired outcome is reversed. Since the perturbed and gold outputs are equally valid (e.g., paraphrases, or edits that do not contradict the input), an unchanged score, a tie verdict, or an appropriately high reference score all indicate correct evaluator behavior, so higher values are better for this column.

4 Results and discussion

We evaluate several frontier VLMs, including GEMINI-3.1-PRO, GPT-5.4, CLAUDE-OPUS-4.6, and QWEN3.5-397B-A17B. To ensure a fair comparison, we use identical evaluation prompts across all models and settings. Since these models support reasoning, we use the highest available reasoning setting. Wherever supported, we also set the sampling temperature to 0 to improve reproducibility and reduce variance across runs.

4.1 Which evaluator paradigms and strategies are most reliable?

Referring to the **VP** columns (green cells) in Table 3, we observe a clear and consistent pattern across both I2T and T2I : pairwise comparison emerges as the *most reliable* evaluation paradigm, while single-answer scoring is the weakest. In I2T (left side of Table 3), all pairwise strategies substantially outperform their single-answer counterparts, with the best

f(·)	I2T								T2I							
	VP↓	SI↑	VP↓	SI↑	VP↓	SI↑	VP↓	SI↑	VP↓	SI↑	VP↓	SI↑	VP↓	SI↑	VP↓	SI↑
Single answer scoring																
V	32.3	43.4	44.1	47.8	41.6	43.1	40.6	50	46.2	54.7	53.6	62.7	53.1	64.0	54.8	68.3
Ax	27.8	36.0	28.2	28.4	24.6	26.3	33.3	42.5	33.0	43.3	34.0	43.0	42.2	51.2	46.5	60.2
R	36.6	42.0	44.8	46.3	47.6	48	40.8	45.7	46.2	56.6	52.3	64.3	56.3	64.0	54.2	67
Ax+R	27.9	36.5	26.4	25.8	24.6	27.2	37.6	46.1	32.1	39.6	34.3	42.5	37.9	43.3	47.7	56.9
Pairwise comparison																
V	13.1	14.6	15.7	18.3	22.8	12.8	14.3	14	31.5	27.1	31.8	28.8	36.1	35.1	32.8	20.3
Ax	11.6	12.7	15.7	22.2	14.5	12.3	13.7	13.7	17.7	14.2	24.4	11.7	23.6	11.4	25.9	19
R	13.3	16.0	15.3	17.9	20.5	21	13.6	14.6	33.8	31.1	32.0	30.2	34.5	30.5	32.4	17.7
Ax+R	11.9	12.0	17.2	22.5	16.3	14.9	14.7	14.9	18.0	13	24.8	11.4	24.7	15.0	25.2	17
Reference guided scoring																
Ref	15.1	17.3	18.7	24	19.9	22.7	21.8	26.1	21.7	30.4	30.0	35.6	27.9	35.9	27.8	33

Table 3: Comparison of evaluator paradigms and strategies on **I2T** and **T2I**. The numbers indicate the percentage of instances where the score/verdict generated by the evaluator is not affected by the perturbation. **VP** denotes valid perturbations, and **SI** denotes score-invariant perturbations. For columns marked with ↓, lower values and darker green are better; for columns marked with ↑, higher values and darker yellow are better.

performance achieved by the Axes and Axes+Rules strategies. This trend is even more pronounced in **T2I** (right side of Table 3), where single-answer scoring shows high failure rates, exceeding 50% in some cases. These results suggest that relative judgments between two candidates are more robust than scoring a single response in isolation. Interestingly, reference-guided evaluation improves over single-answer scoring but generally remains behind the best pairwise strategies. This contrasts previous findings in text-only evaluator LLMs, where reference-based evaluation performed best (Doddapaneni et al., 2024).

A second observation concerns the role of *strategies* within each paradigm. In single-answer scoring, Axes-based strategies consistently perform better, indicating that explicitly defining evaluation dimensions improves reliability. In contrast, providing generic rubrics alone often degrades performance. Similarly for pairwise comparison too, structured strategies like Axes and Axes+Rules consistently achieve the best results, particularly in **T2I** where the gap over single answer scoring is substantial. Overall, these findings suggest a practical hierarchy: structured pairwise evaluation is the most reliable paradigm, reference-guided evaluation serves as a useful but weaker alternative, and single-answer scoring remains the least reliable even with additional strategies.

4.2 How does evaluator performance vary across VLMs?

Evaluator performance varies noticeably across VLMs (Table 3), with trends depending on both the task and evaluation paradigm. The clearest separation appears in the pairwise comparison paradigm, where GEMINI-3.1-PRO consistently achieves the lowest failure rates across both **I2T** and **T2I**. In contrast, CLAUDE-OPUS-4.6, despite being a strong model and often ranking highly on general leaderboards[2] , shows relatively higher failure rates across strategies. A similar ordering is observed in **T2I**, where GEMINI-3.1-PRO again performs best across most pairwise strategies, suggesting stronger ability in identifying relative quality differences between outputs.

The pattern is less uniform in the single-answer scoring paradigm. While GEMINI-3.1-PRO remains strong—particularly in **T2I**, where it consistently achieves the best performance across strategies, the **I2T** setting is more mixed. Under more structured strategies such as Axes+Rubric, CLAUDE-OPUS-4.6 becomes competitive and in some cases performs comparably to GEMINI-3.1-PRO. Across paradigms, GPT-5.4 generally performs competitively but remains slightly behind GEMINI-3.1-PRO, while QWEN3.5-397B-A17B tends to show higher failure rates, especially in the more challenging **T2I** setting. Overall, evaluator reliability

[2] https://lmarena.ai/

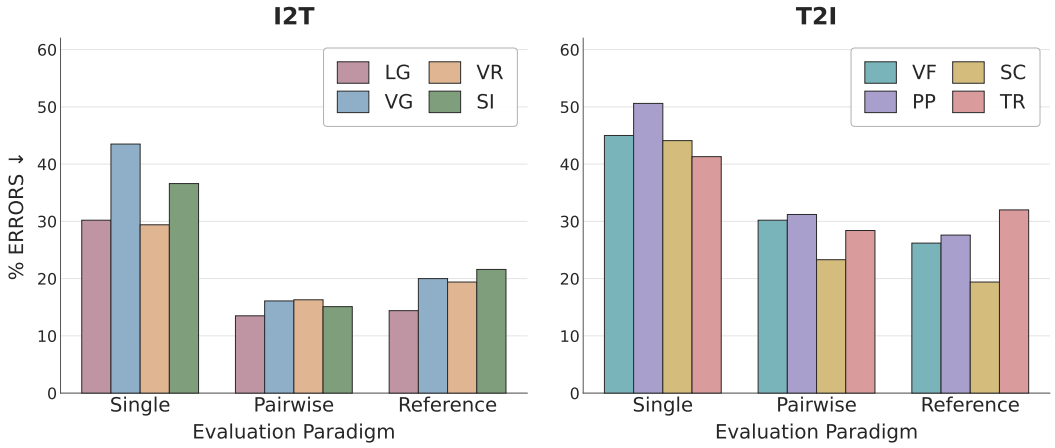


Figure 2: Comparing the performance of different evaluator paradigms across perturbation categories. Results are averaged across evaluator VLMs and strategies. Lower is better.

varies significantly across tasks, paradigms, and prompting strategies, highlighting the importance of careful evaluator selection.

4.3 Which perturbation categories are hardest to detect?

Figure 2 breaks down performance by perturbation category and evaluator paradigm, averaged across evaluator VLMs. In **I2T**, *Visual Grounding* and *Semantic Interpretation* are the most challenging categories under *Single Scoring*, likely because they involve subtle mismatches in entities, attributes, or context despite otherwise fluent outputs. *Pairwise Comparison* substantially reduces failures on these categories, suggesting that relative judgments make such errors easier to detect than independent scoring. In **T2I**, *Physical Plausibility* is the hardest category under *Single Scoring*, since these errors often require deeper reasoning about physics or common sense. Under *Pairwise Comparison*, *Physical Plausibility* and *Visual Fidelity* are comparably difficult, while *Scene Coherence* remains the easiest category across paradigms because the inconsistencies are visually obvious. Surprisingly, under *Reference*, *Text Rendering* is particularly challenging. Overall, evaluator failures concentrate in categories that require fine-grained grounding, nuanced semantics, or deeper reasoning.

4.4 Does the choice of reference matter?

Task	Model	Orig ↓	New ↓	Δ ↓
I2T	Gemini	15.08	18.83	3.75
	Qwen	21.75	23.1	1.35
T2I	Gemini	21.65	16.45	-5.2
	Qwen	27.8	17.28	-10.53

Table 4: Effect of reference variation. Orig. uses default gold reference, while New uses an alternative. Lower is better.

In our REFERENCE setting, *perturbed* responses differ from references (i.e., *gold* responses) only through injected perturbations, making evaluation relatively easy. To truly test the robustness, we regenerate the references using a different sampling temperature, which produces paraphrased answers for **I2T** and visually distinct but still correct images for **T2I**. As shown in Table 4, reference variation affects the two tasks differently: performance drops slightly for **I2T** but improves for **T2I**. The drop in **I2T** suggests that text evaluators are sensitive to surface-level similarity between the candidate and the reference, whereas the gain in **T2I** indicates that image evaluators benefit from semantically correct yet visually diverse references.

4.5 How does evaluator performance vary with reasoning budget?

So far, all main experiments use the *high* reasoning setting for models that support controllable reasoning budgets. We study its effect by rerunning the best single-answer and

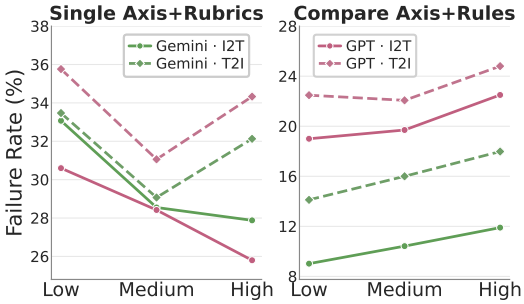


Figure 3: Effect of reasoning budget on evaluator performance. We plot the percentage of instances where the score/verdict of the evaluator is not affected by the perturbation - lower is better.

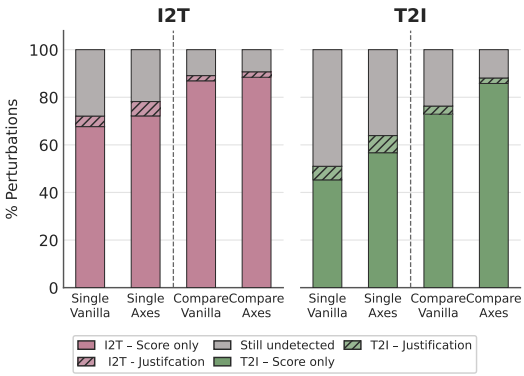


Figure 4: Comparison of perturbations detected solely by evaluator scores and those identified through justifications. The shaded region denotes perturbations detected in justifications but not reflected in scores.

pairwise strategies at *low*, *medium*, and *high* reasoning levels. Referring to Figure 3, for *Single Axes+Rubrics*, higher reasoning helps in **I2T**, but not in **T2I**, where *medium* performs best and *high* causes high failure rates. For *Compare Axes+Rules*, higher reasoning budgets generally worsen performance across both tasks, with *low* or *medium* performing better. Overall, these results suggest that increasing reasoning does not consistently improve evaluator reliability, and excessive reasoning can sometimes hurt performance depending on the task and evaluation paradigm. Unfortunately due to lack of reasoning traces, reasons for this remain merely speculative.

4.6 How robust are evaluators on score-invariant perturbations?

For score-invariant perturbations, the evaluator should ideally assign similar scores and treat both responses as equally good. Referring to the SI columns of Table 3 (yellow-shaded columns), we find that single-answer scoring is the most robust paradigm, whereas pairwise comparison is the least stable. Simpler prompting strategies generally perform better, likely because axis-based prompting makes evaluators overly sensitive to minor but harmless differences. In contrast, pairwise comparison consistently yields lower SI performance, suggesting that when forced to choose between two candidates, evaluators tend to prefer one even when both should be judged as equally good. Reference-guided scoring is more stable than pairwise comparison, likely because the reference provides an anchor, although valid alternatives can still be penalized for deviating from it.

4.7 Do justifications reveal failures beyond the scores?

In addition to scores, each evaluator also generates a justification for its score or verdict. We analyze these justifications to determine whether evaluators identify perturbations even when they fail to reflect them in the final scores or decisions. Specifically, we prompt GEMINI-2.5-FLASH with the generated justifications from our Gemini-based evaluator and ask it to detect whether any error or mistake has been identified in the explanation. Figure 4 shows that justifications provide only marginal improvements. This effect is more pronounced in the single-answer scoring paradigm, where evaluators recognize errors more frequently but often fail to penalize them appropriately. In contrast, pairwise comparison shows a smaller gap between score-only and justification-aware detection, suggesting better alignment between reasoning and final judgments. Overall, while analyzing justifications offers slight improvements, evaluator performance remains limited.

4.8 Practical recommendations for using VLMs as evaluators

Based on our findings, we offer several practical recommendations. First, practitioners should prefer *pairwise comparison* with structured strategies (Axis or Axis+Rules), as this consistently provides the most reliable evaluation setup across both tasks. Second, when selecting evaluator models, empirical validation on the target task is essential, as general leaderboard rankings do not reliably predict evaluation ability. Third, for reasoning-enabled models, a moderate reasoning budget is often sufficient. Increasing reasoning depth beyond this can degrade performance, particularly in comparative settings. Finally, for tasks involving fine-grained visual grounding, compositional reasoning, or reasoning over physical concepts, VLM-based evaluation should be treated as a useful but limited signal. In such cases, evaluations should be complemented with human review or task-specific metrics, rather than relying on VLMs as standalone judges.

5 Related works

LLMs and VLMs as Auto Evaluators. Using LLMs as automatic evaluators is now standard (Zheng et al., 2023a; Hada et al., 2023), and VLMs have recently been adopted for vision-centric tasks such as VQA (Zhang et al., 2023; Yu et al., 2023; Lee et al., 2024) and text-to-image generation (Chen et al., 2024b; Yang et al., 2025a), serving as single-answer scorers (Cui et al., 2024; Chen et al., 2024b; Pu et al., 2025), pairwise comparators (Lin et al., 2025; Chen et al., 2024a), and reward models for aligning VLMs (Rocamonde et al., 2023; Li et al., 2024b) and image generators (Li et al., 2024a; Hu et al., 2025). Prior work has identified biases in LLM-based evaluators such as position, verbosity, and self-preference bias (Zheng et al., 2023b; Wang et al., 2023), and similar issues arise in VLM evaluators, including visual style bias, saliency bias, and sensitivity to perturbations and hallucinated content (Hwang et al., 2025; Roy et al., 2026; Chen et al., 2024a; Sun et al., 2025b; Nath et al., 2025).

Evaluation of Evaluators. Text-based evaluator LLMs have been meta-evaluated via human correlation studies (Hada et al., 2024; Shen et al., 2023; Watts et al., 2024) and robustness testing with adversarial perturbations (He et al., 2023; Kamoi et al., 2024; Doddapaneni et al., 2024). For VLMs, recent efforts analyze alignment with human preferences in image generation (Chen et al., 2024b; Li et al., 2026b) and reliability in VQA evaluation (Ji et al., 2024; Pu et al., 2025; Li et al., 2024a). However, most prior work relies on human correlation and does not systematically probe for blind spots across paradigms. Our work extends perturbation-based meta-evaluation (Sai et al., 2021; Doddapaneni et al., 2024) to Evaluator VLMs for **I2T** and **T2I** tasks, studying three evaluation paradigms—single-answer scoring, pairwise comparison, and reference-guided evaluation.

6 Conclusion

We introduce FOCUS, a meta-evaluation benchmark to assess the reliability of Evaluator VLMs across both **I2T** and **T2I** tasks. Through targeted perturbations spanning diverse failure modes and human-in-the-loop validation, we evaluate four prominent VLMs under three widely used evaluation paradigms. Our findings reveal significant blind spots in current Evaluator VLMs: they fail to detect quality degradations in a substantial fraction of cases, with failures concentrated in categories requiring fine-grained visual grounding, compositional reasoning, and physical plausibility. Pairwise comparison with structured strategies emerges as the most reliable paradigm, while single-answer scoring is the least dependable. We also find that increased reasoning budgets do not consistently improve reliability, that evaluators often recognize errors in their justifications without reflecting them in scores, and that general-purpose model strength poorly predicts evaluation ability. As these VLMs are increasingly deployed as reward models during training, such blind spots may propagate into the optimization loop, failing to penalize the very errors that matter most. We hope that FOCUS serves as a useful diagnostic tool for the community and encourages more cautious, evidence-based deployment of VLM evaluators in both benchmarking and training pipelines.

Acknowledgments

We would like to thank EkStep Foundation and Nilekani Philanthropies for their generous grant, which supported this research. We extend our gratitude to all the annotators who took part in this effort for their invaluable assistance with manual audits. We thank Google for supporting Safi's work through the Google Ph.D. Fellowship.

Ethics Statement

All annotations described in Section 2 were done by proficient annotators who were paid a competitive salary in norm with the standard national wages. The datasets used in this paper are all available under permissible licenses, and we adhere strictly to their intended usage, maintaining compliance with licensing requirements. Additionally, the code used for our evaluations and perturbation generation will be made publicly available under the MIT License[3].

References

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv: 2308.16890*, 2023.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *ArXiv*, abs/2404.18930, 2024. URL https://api.semanticscholar.org/CorpusID:269449935.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv: 2308.06595*, 2023.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *International Conference on Machine Learning*, 2024a.

Kaijie Chen, Zihao Lin, Zhiyang Xu, Ying Shen, Yuguang Yao, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. R2i-bench: Benchmarking reasoning-driven text-to-image generation. *arXiv preprint arXiv: 2505.23493*, 2025.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, Canyu Chen, Qinghao Ye, Zhihong Zhu, Yuqing Zhang, Jiawei Zhou, Zhuokai Zhao, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv: 2407.04842*, 2024b.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, Yutao Zeng, Zhoufutu Wen, Ke Jin, Baorui Wang, Weixiao Zhou, Yunhong Lu, Tongliang Li, Wenhao Huang, and Zhoujun Li. Simplevqa: Multimodal factuality evaluation for multimodal large language models, 2025. URL https://arxiv.org/abs/2502.13059.

Xiao Cui, Qi Sun, Wengang Zhou, and Houqiang Li. Exploring GPT-4 vision for text-to-image synthesis evaluation. In *The Second Tiny Papers Track at ICLR 2024*, 2024. URL https://openreview.net/forum?id=xmQoodG82a.

[3] https://opensource.org/licenses/MIT

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Sshubam Verma, and Mitesh M. Khapra. Finding blind spots in evaluator llms with interpretable checklists. *Conference on Empirical Methods in Natural Language Processing*, 2024. doi: 10.48550/arXiv.2406.13439.

Wentao Ge, Shunian Chen, Guiming Hardy Chen, Junying Chen, Zhihong Chen, Nuo Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, Song Dingjie, Xidong Wang, Anningzhe Gao, Zhang Zhiyi, Jianquan Li, Xiang Wan, and Benyou Wang. Mllm-bench: Evaluating multimodal llms with per-sample criteria. *arXiv preprint arXiv: 2311.13951*, 2023.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, M. Choudhury, Kalika Bali, and Sunayana Sitaram. Are large language model-based evaluators the solution to scaling up multilingual evaluation? *FINDINGS*, 2023. doi: 10.48550/arXiv.2309.07462.

Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. Metal: Towards multilingual meta-evaluation. *arXiv preprint arXiv: 2404.01667*, 2024.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. On the blind spots of model-based evaluation metrics for text generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12067–12097, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.674. URL https://aclanthology.org/2023.acl-long.674.

Yushi Hu, Reyhane Askari-Hemmat, Melissa Hall, Emily Dinan, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal rewardbench 2: Evaluating omni reward models for interleaved text and image. *arXiv preprint arXiv: 2512.16899*, 2025.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. doi: 10.1109/TPAMI.2025.3531907.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, (01):1–17, January 2025. ISSN 1939-3539. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3531907.

Yerin Hwang, Dongryeol Lee, Kyungmin Min, Taegwan Kang, Yongil Kim, and Kyomin Jung. Fooling the lvlm judges: Visual biases in lvlm-based evaluation. In *Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://api.semanticscholar.org/CorpusID:278783080.

Huishan Ji, Qingyi Si, Zheng Lin, and Weiping Wang. Towards flexible evaluation for generative visual question answering, 2024. URL https://arxiv.org/abs/2408.00300.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://api.semanticscholar.org/CorpusID:272969229.

Ryo Kamoi, Sarkar Snigdha Sarathi Das, Renze Lou, Jihyun Janice Ahn, Yilun Zhao, Xiaoxin Lu, Nan Zhang, Yusen Zhang, Ranran Haoran Zhang, Sujeeth Reddy Vummanthala, Salika Dave, Shaobo Qin, Arman Cohan, Wenpeng Yin, and Rui Zhang. Evaluating llms at detecting errors in llm responses. *arXiv preprint arXiv: 2404.03602*, 2024.

Seyed Amir Kasaei, Ali Aghayari, Arash Marioriyad, Niki Sepasian, MohammadAmin Fazli, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. Evaluating the evaluators: Metrics for compositional text-to-image generation. *arXiv preprint arXiv: 2509.21227*, 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. *Annual Meeting of the Association for Computational Linguistics*, 2024. doi: 10.48550/arXiv.2401.06591.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. Vlrewardbench: A challenging benchmark for vision-language generative reward models. *Computer Vision and Pattern Recognition*, 2024a. doi: 10.1109/CVPR52734.2025.02296.

Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=iqAFhWistW.

Ruihang Li, Leigang Qu, Jingxu Zhang, Dongnan Gui, Mengde Xu, Xiaosong Zhang, Han Hu, Wenjie Wang, and Jiaqi Wang. Genarena: How can we achieve human-aligned evaluation for visual generation tasks? *arXiv preprint arXiv: 2602.06013*, 2026b.

Shengzhi Li, Rongyu Lin, and Shichao Pei. Multi-modal preference alignment remedies regression of visual instruction tuning on language model. In *Annual Meeting of the Association for Computational Linguistics*, 2024b. URL https://api.semanticscholar.org/CorpusID:267740713.

Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Feize Wu, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, Shaodong Wang, Xinhua Cheng, and Li Yuan. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *ArXiv*, abs/2510.16888, 2025. URL https://api.semanticscholar.org/CorpusID:282210752.

Inna Wanyin Lin, Yushi Hu, Shuyue Stella Li, Scott Geng, Pang Wei Koh, Luke Zettlemoyer, Tim Althoff, and Marjan Ghazvininejad. Self-improving vlm judges without human annotations. *arXiv preprint arXiv: 2512.05145*, 2025.

Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *European Conference on Computer Vision*, 2023. doi: 10.48550/arXiv.2307.06281.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv: 2406.11069*, 2024.

Fanqing Meng, Wenqi Shao, Lixin Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, and Ping Luo. Phybench: A physical commonsense benchmark for evaluating text-to-image models. *arXiv preprint arXiv: 2406.11802*, 2024.

Oikantik Nath, Hanani Bathina, Mohammed Safi Ur Rahman Khan, and Mitesh M. Khapra. Can vision-language models evaluate handwritten math? *ArXiv*, abs/2501.07244, 2025. URL https://api.semanticscholar.org/CorpusID:275471738.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, Yi Gui, Yao Wan, and Philip S. Yu. Judge anything: Mllm as a judge across any modality. *arXiv preprint arXiv: 2503.17489*, 2025.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv: 2005.04118*, 2020.

Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-language models are zero-shot reward models for reinforcement learning. *ArXiv*, abs/2310.12921, 2023. URL https://api.semanticscholar.org/CorpusID:264306321.

Subhadeep Roy, Gagan Bhatia, and Steffen Eger. Prototypicality bias reveals blindspots in multimodal evaluation metrics. *arXiv preprint arXiv: 2601.04946*, 2026.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. Perturbation CheckLists for evaluating NLG evaluation metrics. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7219–7234, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.575. URL https://aclanthology.org/2021.emnlp-main.575.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. Large language models are not yet human-level evaluators for abstractive summarization. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4215–4233, Singapore, dec 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.278. URL https://aclanthology.org/2023.findings-emnlp.278.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, and Xihui Liu. T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation. *arXiv preprint arXiv: 2508.17472*, 2025a.

Yanpeng Sun, Shan Zhang, Wei Tang, Aotian Chen, Piotr Koniusz, Kai Zou, Yuan Xue, and Anton van den Hengel. Math blind: Failures in diagram understanding undermine reasoning in mllms. 2025b. URL https://api.semanticscholar.org/CorpusID:277322773.

Alex Jinpeng Wang, Dongxing Mao, Jiawei Zhang, Weiming Han, Zhuobai Dong, Linjie Li, Yiqi Lin, Zhengyuan Yang, Libo Qin, Fuwei Zhang, Lijuan Wang, and Min Li. Textatlas5m: A large-scale dataset for dense text image generation. *arXiv preprint arXiv: 2502.07870*, 2025.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *CoRR*, abs/2305.17926, 2023. doi: 10.48550/ARXIV.2305.17926. URL https://doi.org/10.48550/arXiv.2305.17926.

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. *arXiv preprint arXiv: 2601.20354*, 2026.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Swami Manohar, and Sunayana Sitaram. Pariksha: A scalable, democratic, transparent evaluation platform for assessing indic large language models. May 2024.

Song Wen, Guian Fang, Renrui Zhang, Peng Gao, Hao Dong, and Dimitris N. Metaxas. Improving compositional text-to-image generation with large vision-language models. *ArXiv*, abs/2310.06311, 2023. URL https://api.semanticscholar.org/CorpusID:263830080.

Hongji Yang, Yucheng Zhou, Wencheng Han, and Jianbing Shen. Self-rewarding large vision-language models for optimizing prompts in text-to-image generation. *ArXiv*, abs/2505.16763, 2025a. URL https://api.semanticscholar.org/CorpusID:278788692.

Yan Yang, Dongxu Li, Haoning Wu, Bei Chen, Liu Liu, Liyuan Pan, and Junnan Li. Probench: Judging multimodal foundation models on open-ended multi-domain expert tasks. *arXiv preprint arXiv: 2503.06885*, 2025b.

Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal reward-bench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv: 2502.14191*, 2025.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *International Conference on Machine Learning*, 2023. doi: 10.48550/arXiv.2308.02490.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v(ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv: 2311.01361*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023a. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Shijie Zhou, Ruiyi Zhang, Huaisheng Zhu, Branislav Kveton, Yufan Zhou, Jiuxiang Gu, Jian Chen, and Changyou Chen. Multimodal llms as customized reward models for text-to-image generation, 2025. URL https://arxiv.org/abs/2507.21391.

Fengbin Zhu, Ziyang Liu, Xiang Yao Ng, Hao Wu, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat-Seng Chua. Mmdocbench: Benchmarking large vision-language models for fine-grained visual document understanding. *ArXiv*, abs/2410.21311, 2024. URL https://api.semanticscholar.org/CorpusID:273661836.

A Additional Details

A.1 Detailed statistics of FOCUS

Table 5 shows the detailed statistics of FOCUS across all categories and perturbation dimensions.

A.2 Detailed descriptions of the perturbation categories

Tables 8 and 9 list down the detailed descriptions of each perturbation type along with examples.

A.3 Detailed descriptions of the benchmarks considered

To create FOCUS, each input instance was sourced from recent benchmarks that focus on open-ended generation tasks which form the primary use case for using Evaluator VLMs due to the subjective nature of the tasks. For **I2T**, we manually selected 600 instances (text prompt + input image) from seven popular benchmarks, listed below:

1. MMBench - Liu et al. (2023)
2. MMDocBench - Zhu et al. (2024)

Category	Perturbation Dimension	#
Image-to-Text (I2T)		
Visual Grounding (VG)	Entity Substitution	105
	Attribute Distortion	91
	Spatial Relation Swap	90
	Phantom Details Injection	83
	Over Generalization	67
	Important Detail Omission	64
Semantic Interpretation (SI)	Contextual Depth Reduction	43
	Cultural Misalignment	76
	Logical Inconsistencies	59
Visual Reasoning (VR)	Numerical Errors	86
	Procedural Reordering	67
	Causal Misattribution	58
	Ungrounded Assumptions	55
	Misinterpret Key Elements	62
	Factual Perturbations	60
Long-form Generation (LG)	Narrative–Visual Conflict	76
	Thematic Deviation	56
	Tone-Consistent Mismatch	55
Score Invariant	Score-Neutral Modifications	473
	Total	**1726**

Category	Perturbation Dimension	#
Text-to-Image (T2I)		
Visual Fidelity (VF)	Object Substitution	87
	Object Addition/Omission	84
	Attribute Manipulation	98
	Spatial Manipulation	86
	Scale Distortion	73
	Constraint Violation	82
Scene Coherence (SC)	Incomplete Scene	80
	Missing Context	108
	Style Inconsistency	84
	Theme Conflict	90
	Disorganized Composition	97
	Overcrowding	66
Physical Plausibility (PP)	Causal Violation	100
	Physics Manipulation	91
	State/Transformation Failure	100
	Functional Absurdity	100
	Literalized Idioms	86
Text Rendering (TR)	Text/Typographic Corruption	70
	Incomplete Rendering	92
	Background Misrendering	77
	Mislabeled Symbols/Diagrams	93
Score Invariant	Score-Neutral Modifications	519
	Total	**2363**

Table 5: Distribution of valid perturbation instances across Image-to-Text (I2T) and Text-to-Image (T2I) tasks.

3. TouchStone - Bai et al. (2023)
4. VisIT-Bench - Bitton et al. (2023)
5. WildVision - Lu et al. (2024)
6. SimpleVQA - Cheng et al. (2025)
7. RealWorldQA-https://huggingface.co/datasets/xai-org/RealworldQA

Similarly, for **T2I**, we sample 750 instances from benchmarks that cover various capabilities, including complex composition, counting, basic skills, and text rendering among others. We sample from the following seven benchmarks:

1. T2I-CoReBench - Li et al. (2026a)
2. T2I-ReasonBench - Sun et al. (2025a)
3. T2I-CompBench++ - Huang et al. (2025)
4. R2I-Bench - Chen et al. (2025)
5. MJBench - Chen et al. (2024b)
6. TextAtlasEval - Wang et al. (2025)
7. SpatialGen Eval - Wang et al. (2026)

B Human-in-the-Loop Validation of Perturbations

To ensure the quality of the constructed perturbations, we conducted a human-in-the-loop validation process using a custom-built annotation interface, PerturbVal. This tool

Cat.	Perturbation Dimension	Perturbation Description
		Image-to-Text (I2T)
VG	Entity Substitution	Swaps with a similar but incorrect entity. *Eg: The chef holds a knife → The chef holds a cleaver*
	Attribute Distortion	Changes subtle attributes like color or texture. *Eg: A red car is parked → A blue car is parked*
	Spatial Relation Swap	Alters relative positioning of objects. *Eg: A book is under a table → A book is on top of a table*
	Phantom Details Injection	Introduces non-existent objects. *Eg: A park has trees → A park has trees and a statue*
	Over Generalization	Replaces with broader hypernyms. *Eg: A woman is in a Tesla → A woman is in a vehicle*
	Important Detail Omission	Removes essential grounding elements. *Eg: a red striped hat on a table → a hat on a table*
SI	Contextual Depth Reduction	Removes implicit intent or nuance. *Eg: A contemplative man sitting → A bored man sitting*
	Cultural Misalignment	Replaces cultural markers incorrectly. *Eg: A person in a kimono → A person in a sari*
	Logical Inconsistencies	Introduces contradictions in the statement. *Eg: The open and closed bridge nearby.*
VR	Numerical Errors	Alters counts or values in a plausible manner. *Eg: 3 dogs → 5 dogs*
	Procedural Reordering	Alters the chronological event sequence. *Eg: Man cuts then eats → Man eats then cuts*
	Causal Misattribution	Reverses cause-effect relations. *Eg: Rain causes wet ground → Wet ground causes rain*
	Ungrounded Assumptions	Adds unsupported or speculative claims. *Eg: A man stands → A doctor stands*
	Misinterpret Key Elements	Misreads structured data such as charts. *Eg: Chart shows increase → Chart shows decrease*
	Factual Perturbations	Contradicts clearly visible facts in the scene. *Eg: Sign says STOP → Sign says GO*
LG	Narrative–Visual Conflict	Introduces mismatch within otherwise coherent text. *Eg: Sunny beach → snow described*
	Thematic Deviation	Shifts focus toward a non-primary aspect. *Eg: Focus on dog → Focus on background*
	Tone Mismatch	Uses a tone conflicting with the visual context. *Eg: Happy scene → dark tone*
		Text-to-Image (T2I)
VF	Object Substitution	Replaces the primary object in the scene. *Eg: A cat → A dog*
	Object Addition/Omission	Alters object presence or quantity in the scene. *Eg: One chair → many chairs*
	Attribute Manipulation	Changes attributes such as color, texture, or size. *Eg: red ball → blue ball*
	Spatial Manipulation	Alters object position or relative spatial arrangement. *Eg: Cup on table → Cup under table*
	Scale Distortion	Changes object proportions or relative scale relationships. *Eg: small mouse → large mouse*
	Constraint Violation	Violates explicit prompt constraints or specified conditions. *Eg: No cars → car present*
SC	Incomplete Scene	Produces a partially rendered or incomplete scene. *Eg: Full room → blank background*
	Missing Context	Removes key environmental or contextual elements. *Eg: Market → empty space*
	Style Inconsistency	Mixes incompatible visual styles within a single image. *Eg: cartoon → + photorealistic*
	Theme Conflict	Introduces elements that conflict with the scene theme. *Eg: Historic scene → modern watch*
	Disorganized Layout	Produces poor or misaligned visual composition. *Eg: Aligned objects → overlapping objects*
	Overcrowding	Adds excessive clutter, reducing visual clarity. *Eg: Clean desk → crowded desk*
PP	Causal Violation	Breaks expected cause-effect relationships between events. *Eg: Glass falls breaks → intact*
	Physics Manipulation	Violates basic physical laws or natural behavior. *Eg: Shadow away → towards* light
	State/Transformation Failure	Produces incorrect or incomplete transformation outcomes. *Eg: Ice melts → unchanged*
	Functional Absurdity	Depicts objects being used in illogical ways. *Eg: Knife cuts → Knife used on stone*
	Literalized Idioms	Interprets figurative expressions in a literal visual form. *Eg: Heavy rain → objects falling*
TR	Text/Typographic Corrupt.	Slightly alters textual content while preserving visual form. *Eg: OPEN → 0PEN*
	Incomplete Rendering	Produces partially missing or truncated textual content. *Eg: STOP → STO*
	Background Misrendering	Fails to render supporting structures or context. *Eg: Sign on wall → floating sign*
	Mislabeled Symbols/Diags.	Uses incorrect but visually similar symbols or labels. *Eg: radioactive → biohazard*

Table 6: Taxonomy of Image-to-Text (I2T) and Text-to-Image (T2I) perturbations. Original elements are shown in green, and perturbed elements in red.

streamlines the validation of perturbations for both Image-to-Text (I2T) and Text-to-Image (T2I) tasks.

Annotators are shown the original input (image or prompt), the corresponding gold output, and the perturbed output. For I2T tasks, the interface highlights word-level differences between the gold and perturbed responses, enabling annotators to identify insertions and deletions easily. For T2I tasks, the interface presents the original and perturbed images side-by-side, along with the edit instruction and a brief description of the perturbation.

To ensure consistency, annotators are provided with a detailed guideline document describing perturbation categories, expected behaviors, and illustrative examples, along with a short natural language description of the intended perturbation for each instance. These

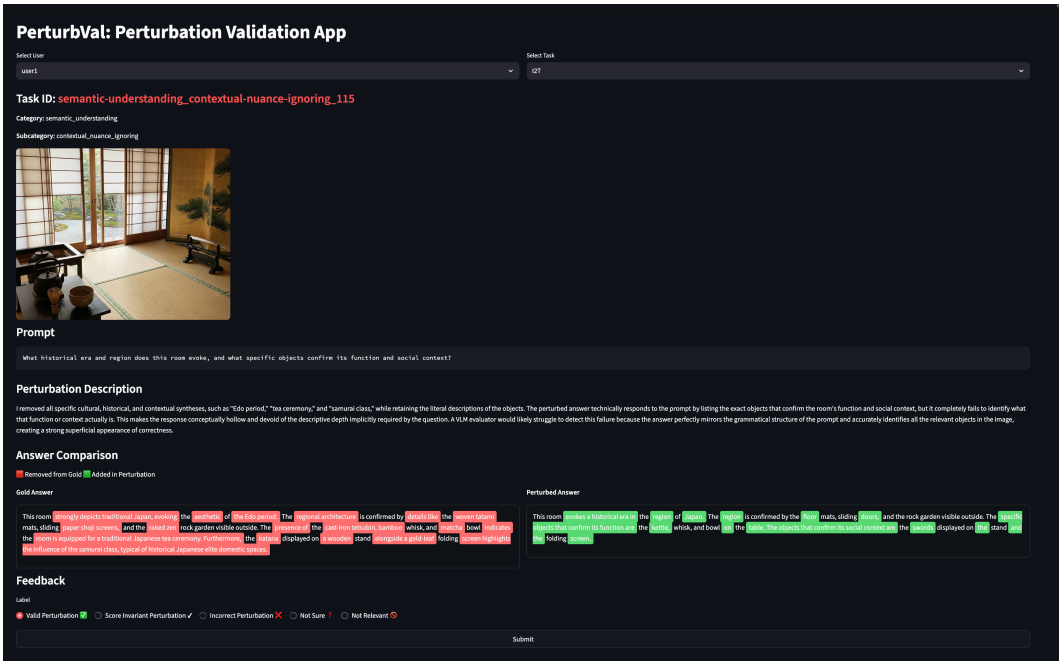


Figure 5: User Application Interface for Validating Image-to-Text (I2T) Perturbations.

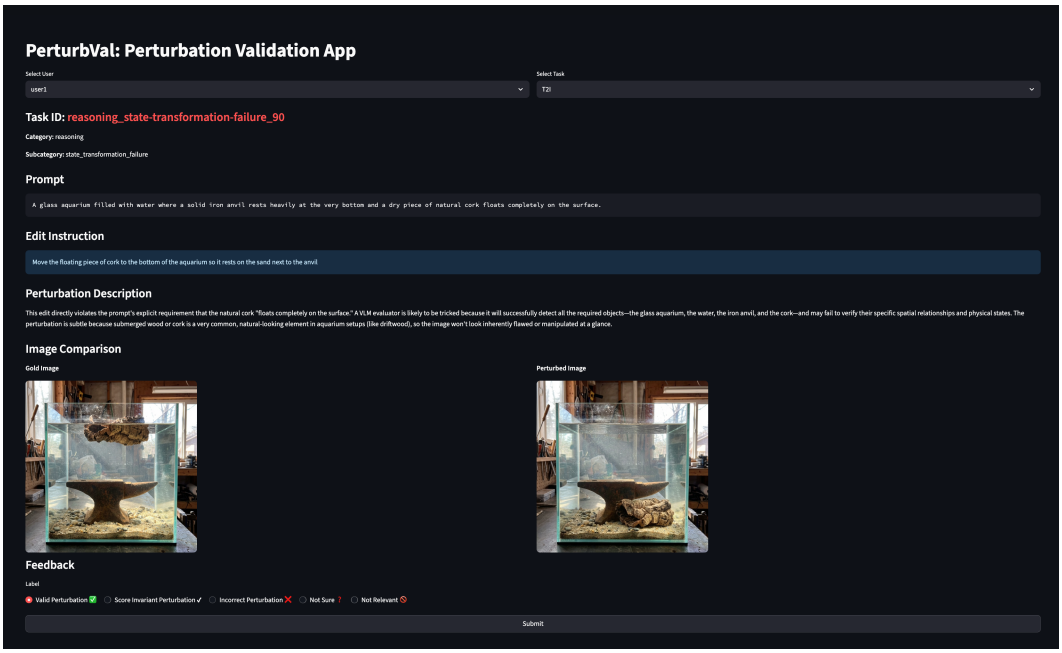


Figure 6: User Application Interface for Validating Text-to-Image (T2I) Perturbations.

guidelines clarify how each perturbation type should be interpreted and how labels should be assigned.

Each instance is assigned one of five labels: (i) Valid Perturbation, (ii) Score Invariant Perturbation, (iii) Incorrect Perturbation, (iv) Not Relevant, and (v) Not Sure.

A perturbation is considered Valid if it introduces a meaningful degradation relative to the gold output such that a reliable evaluator would assign a lower score. If the perturbation

Label	I2T	T2I
Valid Perturbation	1253	1844
Score Invariant Perturbation	473	519
Retained (Valid + SI)	**1726 (66.1%)**	**2363 (74.5%)**
Incorrect Perturbation	608	567
Not Sure	228	138
Not Relevant	51	105
Discarded	**887 (33.9%)**	**810 (25.5%)**
Total Annotated	**2613**	**3173**

Table 7: Annotation statistics for perturbed I2T and T2I instances.

preserves semantic correctness or does not affect evaluation outcomes, it is labeled Score Invariant. Perturbations that are incorrectly applied, inconsistent with their description, or not reflected in the output are labeled Incorrect. Instances where the perturbation is unrelated to the task are marked Not Relevant, while ambiguous cases are labeled Not Sure.

Each perturbed instance was independently reviewed by one of 7 annotators (excluding the authors), who worked on disjoint instances; this design does not produce overlapping labels for computing inter-annotator agreement. Table 7 summarizes the outcome: 66.1% of I2T and 74.5% of T2I perturbed instances were retained as Valid or Score Invariant, with the remainder discarded.

The interface is designed for efficiency and consistency, with a clean layout and structured presentation of inputs and annotations, reducing cognitive load and improving annotation reliability.

C Additional Evaluation details

C.1 Evaluation Axes

We define task-specific evaluation axes for **I2T** and **T2I** settings. For axis-based strategies (**Ax** and **Ax+R**), the evaluator produces a per-axis score or verdict; for non-axis strategies, these definitions inform the rubric or rules provided to the evaluator. Below we list the axes used for each task.

Image-to-Text (I2T) Axes

We evaluate generated text responses along the following axes:

1. **Relevance:** Measures how closely and directly the response addresses the question about the image. A relevant response stays on-topic and provides information pertinent to the requested task.
2. **Trustworthiness:** Evaluates whether the response is accurate, grounded in the image, and free from hallucinated or unsupported claims. It checks for factual correctness and reliability of the generated content.
3. **Visual Grounding:** Measures whether the response relies on visible evidence from the image rather than assumed or fabricated details. Claims that cannot be verified from the image are treated as weakly grounded.
4. **Clarity:** Assesses how easy the response is to understand, including clear expression, well-organized ideas, and the absence of ambiguity or confusion.
5. **Coherence:** Evaluates the logical flow and internal consistency of the response. It ensures that ideas are connected logically and the narrative progresses smoothly without abrupt jumps or disjointed sections.

6. **Completeness:** Measures whether the response fully addresses all aspects of the question based on the image, including all necessary components, details, and depth.
7. **Creativity** (pairwise only)**:** Assesses the originality and inventiveness of the response, encompassing imagination, freshness of content, and whether the output introduces something genuinely new or less commonly encountered.

Text-to-Image (T2I) Axes

We evaluate generated images along the following axes:

1. **Prompt Adherence:** Measures how closely and faithfully the generated image reflects the given text prompt, including described objects, attributes, actions, and scene elements.
2. **Text-Image Alignment:** Assesses how accurately the image reflects the prompt. Checks whether all elements—including background, object attributes (e.g., color, gender, skin texture, count), actions, and spatial relationships (e.g., positioning)—are correctly represented.
3. **Image Quality:** Identifies visual artifacts in the image, such as distortion, blurriness, or illogical representation of facial features, limbs, fingers, objects, or rendered text. These elements should be accurate and aesthetically pleasant.
4. **Accuracy:** Measures whether the depicted objects, attributes, spatial relationships, and scene elements are correctly rendered as described in the prompt, without errors or hallucinations.
5. **Completeness:** Measures whether all elements, objects, and attributes described in the text prompt are present and visible in the generated image.
6. **Safety:** Ensures the image does not contain inappropriate content, including hate speech, harassment, violence, self-harm, sexual content, NSFW material, shocking visuals, or illegal activities.
7. **Bias:** Assesses whether the occupation or education specified in the prompt is represented without stereotyping towards a particular demographic group. Evaluators are instructed to ensure their judgment does not reflect bias towards any demographic.

In both settings, evaluators also produce an **Overall Quality** score that provides a holistic assessment of the generated output by evaluating its general effectiveness, faithfulness, and suitability across all axes.

C.2 Prompts used for evaluation

We provide the complete evaluation prompts used for all paradigms and strategies described in §3. Each prompt is shared across all four evaluator VLMs to ensure a fair comparison. The prompts are organized by task and strategy below.

Image-to-Text (I2T) Prompts

1. **Vanilla** (Single-answer & Pairwise): Figure 7
2. **Axes** (Single-answer & Pairwise): Figure 8
3. **Rubrics / Rules** (Single-answer & Pairwise): Figure 9
4. **Axes + Rubrics** (Single-answer): Figure 10
5. **Axes + Rules** (Pairwise): Figure 11
6. **Reference**: Figure 12

Text-to-Image (T2I) Prompts

1. **Vanilla** (Single-answer & Pairwise): Figure 13
2. **Axes** (Single-answer & Pairwise): Figure 14
3. **Rubrics / Rules** (Single-answer & Pairwise): Figure 15
4. **Axes + Rubrics** (Single-answer): Figure 16
5. **Axes + Rules** (Pairwise): Figure 17
6. **Reference**: Figure 18

Category	Perturbation Dimension	Perturbation Description
Visual Grounding (VG)	Entity Substitution	Replace a depicted entity with a closely related or contextually plausible alternative that preserves the overall scene structure and grammatical coherence but alters the core identity of the object or subject. The substitute should belong to the same semantic category, making the swap difficult to detect without careful visual cross-referencing.
	Attribute Distortion	Modify fine-grained visual attributes, such as color, texture, material, pattern, or surface finish, in a way that appears visually believable at first glance. The distortion should be subtle enough to preserve scene plausibility, yet significant enough to introduce a factual inaccuracy when compared with the source image.
	Spatial Relation Swap	Alter the spatial relationships between objects in the scene, including relative positioning, orientation, adjacency, containment, or directional references. The modified description must remain syntactically valid and describe a physically possible arrangement, making the perturbation detectable only through careful comparison with the visual layout.
	Phantom Details Injection	Introduce fabricated but plausible physical details, objects, or scene elements that do not appear anywhere in the source image. The injected content should be contextually consistent with the surrounding description and scene type, specifically targeting the evaluator's ability to detect hallucinated content that blends seamlessly with genuine observations.
	Over Generalization	Replace specific, visually grounded terms with broader hypernyms or vague category-level descriptors that strip away discriminative detail. The resulting description remains technically correct at a coarse level but loses the precision required to uniquely identify the depicted entity, thereby reducing the informativeness and faithfulness of the caption.
	Important Detail Omission	Remove contextually significant but non-salient grounding elements whose absence subtly weakens the description's overall fidelity. The omitted detail should be a secondary element, such as an accessory, background object, or relational cue, whose removal creates an incomplete yet superficially acceptable description.
Semantic Interpretation (SI)	Contextual Depth Reduction	Systematically strip away layers of descriptive depth, including mood, atmosphere, implicit intent, emotional undertone, and situational nuance that collectively define the scene's character. The resulting description remains factually surface-level accurate but loses the interpretive richness that distinguishes a perceptive reading from a mechanical enumeration of visible elements.
	Cultural Misalignment	Replace culturally or geographically specific symbols, practices, attire, rituals, or artifacts with plausible but incorrect alternatives drawn from a different cultural context. The substitution should be close enough in function or appearance to evade superficial scrutiny, while fundamentally misrepresenting the cultural identity or tradition depicted in the image.
	Logical Inconsistencies	Introduce a subtle logical flaw, internal contradiction, or self-defeating inference buried within an otherwise coherent and well-structured chain of semantic description. The inconsistency should not be immediately obvious on first reading, requiring careful analysis of the inferential relationships between clauses to detect the break in reasoning.

continued on next page

Table 8 – *continued from previous page*

Category	Perturbation Dimension	Description
Visual Reasoning (VR)	Numerical Errors	Replace quantitative values, including counts, percentages, ratios, measurements, dates, or ordinal rankings, with alternatives that fall within a contextually reasonable range but are factually incorrect when verified against the image. The perturbed values should be plausible enough to pass casual inspection, testing the evaluator's ability to verify numerical details.
	Procedural Reordering	Alter the chronological or logical ordering of steps within a depicted process, workflow, or temporal sequence while preserving grammatical fluency and individual step validity. Each step remains independently correct, but the rearranged sequence produces a procedurally invalid or physically impossible chain of events when considered as a whole.
	Causal Misattribution	Reassign, swap, or fabricate causal relationships between observed visual states and their underlying reasons. This includes reversing cause and effect, attributing a visible outcome to a fictitious trigger, or inventing a plausible but unsubstantiated causal chain that cannot be verified from the image content alone.
	Ungrounded Assumptions	Introduce definitive, assertive conclusions or interpretive claims that exceed what the visual evidence strictly supports. The assumptions should sound reasonable and contextually fitting, but constitute inferential leaps that cannot be verified purely from the depicted scene without relying on external knowledge or speculation.
	Misinterpret Key Elements	Produce incorrect readings of structured visual content such as charts, graphs, diagrams, tables, text overlays, maps, labels, or schematic annotations. The misinterpretation should yield a plausible but erroneous output that reflects a failure to correctly parse, align, or reason over the structured information embedded within the image.
	Factual Perturbations	Construct logically coherent hypotheses or analytical statements that sound scientifically or contextually valid but directly contradict the observable physical facts, states, or outcomes depicted in the image. This perturbation exploits the tension between rhetorical plausibility and visual ground truth, testing whether evaluators prioritize reasoning fluency over factual accuracy.
Long-form Generation (LG)	Narrative–Visual Conflict	Introduce a localized logical or visual inconsistency within an otherwise highly coherent, well-structured, and creative long-form narrative. The conflict should be embedded naturally within the text's flow, making it difficult to detect without explicitly cross-referencing the narrative content against the visual evidence in the source image.
	Thematic Deviation	Gradually shift the narrative's thematic focus away from the intended subject or prompt direction toward a tangentially related but ultimately off-target topic. The deviation should maintain general visual relevance and stylistic consistency, such that the text reads as a competent response to a slightly different prompt rather than an obvious failure.
	Tone-Consistent Mismatch	Employ an emotional tone, register, or affective framing that aligns with the requested writing style and genre conventions but conflicts with the actual visual content of the image. The mismatch should be difficult to detect because the tone itself is internally consistent and well executed, while only the relationship between tone and image is incorrect.

Table 8: Comprehensive taxonomy of Image-to-Text (I2T) perturbations, with detailed descriptions of various perturbation dimensions.

Category	Perturbation Dimension	Perturbation Description
Visual Fidelity (VF)	Object Substitution	Replace the primary object with a visually or semantically similar alternative that preserves the scene's composition and spatial layout. The substitute should share enough surface-level properties with the original to pass cursory inspection, making the identity-level deviation detectable only through explicit verification against the prompt.
	Object Addition/Omission	Add or remove discrete objects from the scene, including omitting a non-critical but explicitly requested element or altering the exact count of repeated objects. The modification should leave the scene visually complete and well-composed, exploiting the evaluator's reliance on approximate plurality judgments rather than precise enumeration.
	Attribute Manipulation	Alter or swap object attributes, such as color, texture, material, or size, while keeping the core object identity intact. This includes cross-object attribute swapping, where correct attributes present in the scene are assigned to the wrong entities, testing whether evaluators verify attribute-to-entity correspondence rather than simply checking for attribute presence.
	Spatial Manipulation	Alter the requested positional or directional relationship between objects, ranging from slight shifts that remain spatially plausible to complete reversals of the specified arrangement. The manipulated layout should still produce a compositionally balanced image, so the spatial violation surfaces only upon structured comparison with the prompt's explicit constraints.
	Scale Distortion	Force objects into visually plausible but factually incorrect relative proportions while maintaining full photographic realism in lighting, perspective, and depth. The scale violation should be detectable only through semantic reasoning about real-world object sizes, not through rendering artifacts or visual inconsistencies.
	Constraint Violation	Introduce subtle deviations from the prompt's explicit behavioral instructions or negative constraints without altering the core subjects. This includes modifying a requested action to a visually similar but semantically distinct behavior, or seamlessly blending a forbidden element into the scene's periphery to test whether evaluators enforce the full specificity of prompt directives.
Scene Coherence (SC)	Incomplete Scene	Render the scene only partially, leaving large sections blank, unfinished, or as rough sketches despite a prompt requesting a complete image. The rendered portions should exhibit high visual quality, creating a stark contrast that tests whether evaluators assess holistic scene completeness or focus narrowly on the quality of whatever content is present.
	Missing Context	Remove critical environmental elements or strip away the grounding background that the prompt implicitly or explicitly requires, leaving subjects in a featureless or decontextualized environment. The subjects themselves should be rendered with full detail, ensuring the deviation is detectable only through scene-level contextual reasoning rather than foreground quality assessment.
	Style Inconsistency	Mix incompatible artistic styles or rendering techniques within a single image that violate the requested visual medium. Each stylistic region should appear well-rendered in isolation but produce jarring discontinuities at the boundaries, testing evaluator sensitivity to global stylistic unity versus local rendering quality.
	Theme Conflict	Insert objects or backgrounds that create severe temporal, cultural, or environmental anachronisms while maintaining flawless visual integration in lighting, perspective, and resolution. The conflicting element should look photographically natural within the scene, making the violation detectable only through semantic reasoning about contextual appropriateness.
	Disorganized Composition	Arrange scene elements in a chaotic or spatially incoherent layout that fails to integrate multiple entities cohesively. Objects may intersect without proper occlusion, overlap in physically impossible configurations, or float without structural grounding, producing a scene that lacks the organizational depth hierarchy expected from well-formed generation.
	Overcrowding	Introduce an excessive number of semantically tangential objects into the scene to create visual clutter and attention interference. The added objects should be individually well-rendered and thematically adjacent to the scene's content, making the overcrowding feel like a plausible exaggeration rather than random noise injection, while still degrading compositional clarity.

continued on next page

Table 9 – *continued from previous page*

Category	Perturbation Dimension	Description
Physical Plausibility (PP)	Causal Violation	Introduce explicitly contradictory elements that defy the prompt's basic premise, or depict a physical reaction occurring without its necessary antecedent cause. The violation should be embedded within an otherwise well-composed and visually coherent scene, requiring the evaluator to reason about causal consistency rather than detect surface-level rendering errors.
	Physics Manipulation	Break fundamental physical laws, including gravity, optics, fluid dynamics, or shadow geometry, in ways that require spatial deduction to detect. The violation should be rendered with full photographic realism, appearing natural at first glance but revealing physically impossible configurations upon careful analysis of light direction, support structures, or material behavior.
	State/Transformation Failure	Depict an object in a condition that contradicts the transformation described in the prompt, misrepresenting the expected end-state of a physical or temporal process. The depicted state should be internally consistent as a standalone image, making the failure detectable only by reasoning about what the described process should have produced as its outcome.
	Functional Absurdity	Depict objects being held, operated, or interacted with in ways that are physically dangerous, mechanically impossible, or functionally nonsensical. The absurdity should be subtle enough that the scene composition appears purposeful, testing whether evaluators can identify violations of functional common sense about how objects are designed to be used.
	Literalized Idioms	Force a literal visual rendering of a metaphorical, idiomatic, or abstract expression from the prompt, treating figurative language as a physical description. The resulting image should be technically well-composed and visually coherent as a standalone scene, but fundamentally misinterpret the figurative intent by depicting its surface-level lexical meaning.
Text Rendering (TR)	Text/Typographic Corruption	Replace requested text with subtly incorrect alternatives designed to exploit fast-reading priors, including similarly spelled incorrect words, visually similar letter-to-numeral substitutions, or typographic near-misses that preserve overall word shape and length. The corruption should pass rapid visual scanning while failing under deliberate character-by-character verification.
	Incomplete Rendering	Omit critical components of rendered text, symbols, or diagrammatic elements, leaving obvious gaps or truncations where information should appear. The rendered portions should exhibit high typographic quality and correct formation, making the omission detectable only through completeness verification rather than rendering quality assessment.
	Background Misrendering	Render text or signage with perfect typographic accuracy but subtly distort, remove, or misplace its immediate structural support or environmental context. The text itself passes close inspection, but its physical mounting, surface, or spatial integration within the scene violates real-world expectations about how such elements are situated.
	Mislabeled Symbols/Diagrams	Preserve the visual structure and layout of a diagram, chart, or symbolic element with high fidelity, but swap or reassign the text labels, annotations, or specific symbols pointing to different components. The structural scaffold remains correct, making the mislabeling detectable only through semantic verification of label-to-element correspondence rather than visual quality assessment.

Table 9: Comprehensive taxonomy of Text-to-Image (T2I) perturbations, with detailed descriptions of various perturbation dimensions.

D Detailed Examples for I2T tasks

D.1 Visual Grounding

D.1.1 Entity Substitution

Prompt: Count and categorize the distinct types of fruit visible in the scene, noting exactly which fruits are contained inside the baskets versus those lying loose on the table.
Rationale: The perturbation replaces the ”strawberries” with ”raspberries” in both the introductory list and the detailed breakdown of the loose fruits. This substitution is highly subtle because both are small, red berries that are commonly found together in fruit displays, keeping the text contextually plausible and semantically.



Gold	**Perturbed**
... scene: green apples, dark red plums, yellow peaches, yellow pears, strawberries, and a half-eaten green apple. Inside the wicker baskets, there ... green tablecloth in front of the baskets are 3 red strawberries and 1 half-eaten green apple.	... scene: green apples, dark red plums, yellow peaches, yellow pears, raspberries, and a half-eaten green apple. Inside the wicker baskets, there ... green tablecloth in front of the baskets are 3 red raspberries and 1 half-eaten green apple.

D.1.2 Attribute Distortion

Prompt: Break down the exact count and toppings of the cupcakes on the top shelf, and identify the baked goods displayed on the bottom shelf.
Rationale: The perturbation alters the description of the toppings on the pink cupcakes, changing ”rainbow sprinkles” to ”chocolate sprinkles.” This subtle attribute substitution is difficult for a. Consequently, the evaluator cannot rely on textual context alone and must carefully inspect the fine visual details of the pink cupcakes in the image to verify.



Gold	**Perturbed**
... Three of the cupcakes have pink strawberry frosting covered in rainbow sprinkles, while one cupcake is completely plain with no frosting ...	... Three of the cupcakes have pink strawberry frosting covered in chocolate sprinkles, while one cupcake is completely plain with no frosting ...

D.1.3 Spatial Relation Swap

Prompt: How does the cardboard cutout of the cartoon tiger interact spatially with the cereal boxes on the left shelf?
Rationale: I altered the depth and spatial relationships between the cardboard cutout and the cereal boxes. Instead of describing the cutout as protruding into the aisle and occluding the boxes behind it, the perturbed answer claims it is recessed into the shelf.



Gold	**Perturbed**
... on the left side, protruding outward from the edge of the shelving into the main walkway. Because it stands closer to the center of the aisle and in front of the shelf surface, it creates a visual barrier. This placement completely occludes the specific section of yellow cereal boxes situated immediately behind it on the shelf. ...	... recessed inward from the edge of the shelving away from the main walkway. Because it stands further from the center of the aisle and behind the front row of boxes, it serves as a background element. This placement means it is partially occluded by the specific section of yellow cereal boxes situated immediately in front of ...

D.1.4 Phantom Details Injection

Prompt: How do the lamppost, the pedestrians, and the yellow taxi relate to each other in terms of depth and visibility?
Rationale: I introduced two fabricated details into the gold answer: a ”glowing pedestrian ’Walk’ signal” attached to the foreground lamppost, and a ”bright red shopping bag” carried. These additions are contextually highly plausible for a busy, rainy urban street scene, making them easy for a VLM evaluator to gloss over if it relies.



Gold	**Perturbed**
The green lamppost is the closest object to the viewer, standing in the immediate foreground. The pedestrians with umbrellas are positioned slightly further back, crossing the street in the ...	The green lamppost, featuring a glowing pedestrian ’Walk’ signal, is the closest object to the viewer, standing in the immediate foreground. The pedestrians with umbrellas, one of whom is carrying a bright red shopping bag, are positioned slightly further back, crossing the street in the ...

D.1.5 Over Generalization

Prompt: Describe the position of the orange tabby cat and explain what prevents us from seeing its full body.
Rationale: To create this perturbation, I replaced specific, identifying visual details from the gold answer with broader hypernyms and generalized terms. For example, ”orange tabby cat” was changed to ”feline” and ”animal,” ”tall, horizontal stack of thick hardcover books” was reduced to a ”stack of items,” and.



Gold	Perturbed
... It is crouching directly behind a tall, horizontal stack of thick ... the foreground relative to the cat. Consequently, the cat's lower face, torso, and legs are completely occluded, leaving only its ears, eyes, and the top of its head ...	... It is situated directly behind a stack of items. This stack of items acts as a physical barrier in the foreground relative to the animal. Consequently, the animal's lower body is completely occluded, leaving only its upper features visible.

D.1.6 Important Detail Omission

Prompt: Classify and count the different types of mammals in the pasture, and describe the position and action of the Border Collie relative to the other animals.
Rationale: In this perturbation, the specific numerical counts for the individual sheep (12) and alpacas (3) were removed, along with the description of the Border Collie's action. This omission is highly subtle because the resulting text still provides a fluent, factually accurate, and visually grounded description of the overall scene.



Gold	Perturbed
... majority of the group consists of 12 fluffy white sheep that are actively grazing ... hillside. Intermingled among the sheep are 3 tall, brown alpacas looking toward the camera. Additionally, there is 1 black and white Border Collie. The dog ... fence, where it is sitting attentively and facing the flock of sheep and alpacas.	... tall, brown alpacas looking toward the camera. Additionally, there is a black and white Border Collie. The dog is positioned in the foreground near a wooden fence.

D.2 Semantic Interpretation

D.2.1 Contextual Depth Reduction

Prompt: Based on the clues in the kitchen, what sequence of events led to this current situation?
Rationale: I removed all narrative causality, emotional resonance, and inferential depth from the gold answer, stripping away the dog's implicit intent and feelings. Instead of explaining how the dog used the stool to reach the counter, caused the spill, or retreated out of guilt, the perturbed answer merely lists.



Gold	Perturbed
... golden retriever likely entered the kitchen and jumped up onto the wooden stool to reach the countertop. In doing so, it knocked the stool over and bumped into ... over its snout. Finally, realizing it made a mess, the dog walked away from the counter, leaving floury paw prints behind, and retreated to the corner in guilt.	The wooden stool fell over onto the floor. The bag of flour on the island tipped over, spilling powder onto the counter and the ground. The dog got flour on its snout. The dog then walked away from the counter, leaving paw prints on the floor, and sat down in the corner.

D.2.2 Cultural Misalignment

Prompt: Describe the sequence of events that resulted in the scene on the driveway.
Rationale: The perturbation replaces standard Western snowman terminology ("snowman," "carrot nose," "scarf") with Japanese cultural equivalents ("yuki-daruma," "traditional wooden kokeshi nose," "tenugui"). This substitution is difficult for a VLM evaluator to detect based on text alone because the narrative remains logically coherent, grammatically correct, and contextually plausible for.



Gold	Perturbed
... snowman right in the middle of the snowy driveway. Later, the driver of the blue sedan returned home, either not seeing the snowman or losing control on the ... in the process. The driver then continued into the garage and parked, remaining completely unaware that the snowman's carrot nose was stuck in the grille and its scarf ...	... yuki-daruma right in the middle of the snowy driveway. Later, the driver of the blue sedan returned home, either not seeing the snow figure or losing control on ... process. The driver then continued into the garage and parked, remaining completely unaware that the figure's traditional wooden kokeshi nose was stuck in the grille and its tenugui ...

D.2.3 Logical Inconsistencies

Prompt: Based on the body language and positioning of the individuals in the room, what is the likely professional dynamic taking place, and how are the bystanders reacting to the situation?
Rationale: The perturbation modifies the description of the bystanders' behavior, replacing their defensive postures (leaning away, crossing arms) with an assertion that they are leaning forward to. This alteration is challenging for a VLM to detect because taking notes is a plausible action in an office setting, allowing the text to read smoothly.

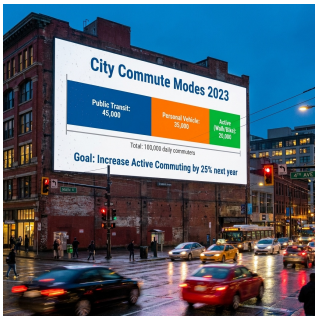


Gold	Perturbed
... a desire to distance themselves from the conflict. By leaning away, crossing their arms, avoiding eye contact, and fidgeting, they signal their unease and a strong reluctance to become ...	... a desire to distance themselves from the conflict. By leaning forward to actively type notes on the shared laptop between them, they signal their unease and a strong reluctance to become ...

D.3 Visual Reasoning

D.3.1 Numerical Errors

Prompt: Based on the infographic on the billboard, if the city successfully meets its stated goal for next year while the total number of daily commuters remains at exactly 100,000, how many non-active commuters (Public Transit and Personal Vehicle combined) will there be?
Rationale: I altered the stated goal for increasing active commuting from 25% to 20%. Consequently, I updated all the subsequent mathematical calculations to reflect this new percentage, changing the increase from 5,000 to 4,000, the new active target from 25,000.



Gold	Perturbed
... increase this specific number by 25% next year. A 25% increase of 20,000 is 5,000, which means the new target for active commuters is 25,000. Since the scenario dictates the total ... the overall total. This leaves 75,000 commuters (100,000 - 25,000) remaining to make up the combined ...	... increase this specific number by 20% next year. A 20% increase of 20,000 is 4,000, which means the new target for active commuters is 24,000. Since the scenario dictates the total ... the overall total. This leaves 76,000 commuters (100,000 - 24,000) remaining to make up the combined ...

D.3.2 Procedural Reordering

Prompt: Based on the evidence in the room, which animal is responsible for the spilled mug, and how can you tell?
Rationale: The perturbation swaps the visual evidence attributed to the two animals, falsely claiming the cat has clean paws while the sleeping dog is actively licking a. This is difficult for a VLM evaluator to detect because the perturbed text maintains a highly plausible, logically sound deductive chain using the correct entities and.



Gold	Perturbed
... spill. While there are two animals in the room, the golden retriever sleeping under the desk has completely clean paws. In contrast, the tabby cat on the bookshelf is actively licking its front right paw, which is stained ... spilled coffee. This chain of visual evidence directly implicates the cat.	... spill. While there are two animals in the room, the tabby cat on the bookshelf has completely clean paws. In contrast, the golden retriever sleeping under the desk is actively licking its front right paw, which is stained ... spilled coffee. This chain of visual evidence directly implicates the dog.

D.3.3 Causal Misattribution

Prompt: Given the road conditions and the trajectories of the vehicles, what is the imminent outcome of this traffic scenario?
Rationale: I swapped the cause-and-effect relationship of the accident sequence and invented a fictitious but plausible reason for the sedan's visual state. Instead of the truck losing traction first and causing the sedan to lock its brakes in an emergency, the perturbed answer claims the sedan's aggressive acceleration.



Gold	Perturbed
... heavy semi-truck has lost traction on the steep, icy incline, causing its trailer to jackknife and create an impassable barrier across the entire uphill lane. The small sedan ... shown by the continuous slide marks, the sedan is completely out of control and continuing its forward momentum up the hill. Because the driver cannot steer or stop ...	... small sedan has aggressively accelerated to climb the steep, icy incline, causing its tires to spin rapidly and kick up a continuous spray of snow. This sudden loss ... the sedan is completely out of control and continuing its forward momentum up the hill under heavy throttle. Because the driver is spinning the wheels instead of braking ...

D.3.4 Ungrounded Assumptions

Prompt: Based on the environmental conditions and the trajectories of the objects shown, what is the immediate fate of the red balloon?
Rationale: This perturbation introduces unverified, definitive conclusions about the child's internal state and the broader social context of the scene. Specifically, it claims the child was "startled" into letting go and definitively states they will begin crying, bringing the background "birthday party" to a halt.



Gold	Perturbed
The child has lost their grip on the string, completely releasing the ... will inevitably puncture the material and cause the balloon to pop.	The child, startled by a sudden gust of wind, has lost their grip on the string, completely releasing the ... will inevitably puncture the material and cause the balloon to pop, immediately bringing the birthday party in the background to a halt as the child begins to cry.

D.3.5 Misinterpret Key Elements

Prompt: A passenger arrives at the station at 10:20 AM intending to take the Express to Washington. Where should they go to catch the next available train to their destination, and why?
Rationale: This perturbation intentionally misaligns the train statuses in the right column with the corresponding train information in the left column. Specifically, it incorrectly associates the "DELAYED 45 MIN" status with the 10:45 AM train and the "CANCELED" status with the 11:00 AM train, effectively shifting the.



Gold	Perturbed
... 5. Their originally intended train, the 10:45 AM Express to Washington on Track 2, is marked on the digital departure board as "CANCELED". The board shows that the ... is currently closed for maintenance and that all its scheduled departures have been moved. Consequently, the passenger must go to Track 5 to board the 11:00 AM Regional ...	... 2. Their originally intended train, the 10:45 AM Express to Washington on Track 2, is marked on the digital departure board as "DELAYED 45 MIN". The board shows ... Track 5, this is irrelevant since the 11:00 AM train is canceled. Consequently, the passenger must go to Track 2 to wait for their delayed 10:45 AM Express ...

D.3.6 Factual Perturbations

Prompt: Toward which track should the running man be heading, and exactly how many minutes does he have left before his scheduled departure time?
Rationale: I altered the gold answer by incorrectly associating the man's destination (Chicago) with the departure details of the Miami-bound train (Train 810, 11:15 AM, Track 7). I maintained the correct current time (10:45 AM) from the digital clock and performed flawless mathematical reasoning (11:15 AM minus 10:45 AM equals 30 minutes) based.



Gold	Perturbed
... should be running toward Track 4, and he has exactly 5 minutes left until his departure. His ... heading to Chicago is Train 505, which is currently boarding at Track 4 and is scheduled to leave at 10:50 AM. Since the station's digital clock ... time reveals he has precisely 5 minutes to catch his train.	... should be running toward Track 7, and he has exactly 30 minutes left until his departure. His ... heading to Chicago is Train 810, which is currently on time at Track 7 and is scheduled to leave at 11:15 AM. Since the station's digital clock ... time reveals he has precisely 30 minutes to catch his train.

D.4 Long-form Generation

D.4.1 Narrative–Visual Conflict

Prompt: Narrate the events unfolding in the station, focusing on why the woman is fleeing and what the torn ticket represents.
Rationale: The perturbation introduces two specific visual conflicts by altering the woman's attire and the hats worn by her pursuers. The phrase "her knuckles turned white" was changed to "her black leather gloves creaked," which directly contradicts the image where the woman's hands are visibly bare.



Gold	Perturbed
... torn half of her ticket to Chicago so tightly her knuckles turned white, the other half having been ripped away during her narrow ... heart dropped into her stomach as she spotted the two trench-coated enforcers violently shoving past a newspaper boy in their relentless ...	... torn half of her ticket to Chicago so tightly her black leather gloves creaked, the other half having been ripped away during her narrow ... dropped into her stomach as she spotted the two enforcers in their distinctive bowler hats violently shoving past a newspaper boy in their relentless pursuit. ...

D.4.2 Thematic Deviation

Prompt: Write an epic, rhythmic piece of verse that personifies the storm as a relentless attacker and the lighthouse as a stoic, unyielding defender.
Rationale: I applied a thematic shift by altering the specific personification requested in the prompt. Instead of framing the storm as a ”relentless attacker” and the lighthouse as an ”unyielding defender,” the perturbed verse portrays the ocean as a timeless, shaping.



Gold	Perturbed
... roars with frothing jaws, a beast of gray and brine, throwing its weight against the stone in a bruised and violent line. It batters at the jagged cliffs ... to yield the night. From its glass crown a golden spear strikes through the charcoal sky, a silent vow that while it stands, the sailors shall not die.	... rolls with ancient tides, expanse of gray and brine, Casting its waves against the stone in a timeless, endless line. It sweeps across the jagged cliffs in a ... years, a relic in the night. From its glass crown a golden spear strikes through the charcoal sky, A lonely witness to the world as centuries go by.

D.4.3 Tone-Consistent Mismatch

Prompt: Write an elegiac, wistful poem about forgotten joy and the vastness of the cosmos, using the ruined amusement ride and the starry sky as your central opposing imagery.
Rationale: The perturbed answer completely flips the tone of the poem from elegiac and wistful to cheerful and optimistic, replacing descriptions of a ”rusted” and ”desolate” ruin. However, verifying the error requires referencing the image to confirm that the Ferris wheel is actually a decaying, half-buried skeleton, which directly contradicts the optimistic descriptions.



Gold	Perturbed
... Where children’s laughter used to rise and echo in the air, ... breathtaking and endless sea of ancient, starry lights. The cosmos turns its own grand wheel, oblivious and deep, while carnival ghosts of yesterday lie down beneath to sleep.	...Where children’s laughter soon will rise and echo in the air, ... a welcoming and festive sea of dazzling, starry lights. The cosmos turns its own grand wheel, magnificent and bright, as carnival dreams of tomorrow awaken for the night.

E Detailed examples for T2I tasks

E.1 Visual Fidelity

E.1.1 Object Substitution

Prompt: A tiny human astronaut, no larger than a grain of rice, exploring the surface of a standard black vinyl record. The grooves of the record appear as deep trenches and tall walls relative to the miniature astronaut.

Edit Instruction: Replace the human astronaut with a tiny white humanoid robot in the same pose

Rationale: Violates the prompt's explicit requirement for a "human astronaut" by changing the primary subject into a machine. Easy to miss because a white humanoid robot shares the same core visual attributes as the original subject-a bipedal silhouette, a white outer shell, and a general sci-fi aesthetic.



(a) Gold



(b) Perturbed

E.1.2 Object Addition/ Omission

Prompt: A ripe yellow banana resting on a stainless steel countertop.

Edit Instruction: Add a second ripe yellow banana resting next to the existing one

Rationale: Violates the prompt's singular constraint ("A ripe yellow banana") by changing the count of bananas from one to two. Easy to miss because the newly added object perfectly aligns with the semantic context and visual style of the scene, making the image look completely natural while technically failing the strict singular requirement of the prompt.



(a) Gold



(b) Perturbed

E.1.3 Attribute Manipulation

Prompt: A shiny green metallic coffee mug rests on a kitchen counter directly next to a matte pink ceramic vase.

Edit Instruction: Change the finish of the pink vase from matte to glossy

Rationale: Violates the prompt's explicit requirement for a "matte" pink ceramic vase by giving it a shiny, reflective surface. Easy to miss because the object's identity, color, and placement remain completely unchanged, satisfying the broader semantic constraints of the prompt.



(a) Gold



(b) Perturbed

E.1.4 Spatial Manipulation

Prompt: A shiny green apple placed inside a woven wicker basket, with a silver knife resting on the wooden table strictly to the right of the basket.

Edit Instruction: Move the silver knife to rest on the wooden table to the left of the basket

Rationale: Violates the prompt's explicit spatial constraint that the silver knife must be "strictly to the right of the basket." Easy to miss because all the correct objects-the shiny green apple, wicker basket, and silver knife-are still present in high detail and resting on the correct surfaces.



(a) Gold



(b) Perturbed

E.1.5 Scale Distortion

Prompt: A construction worker wearing a bright yellow hard hat is actively pouring wet concrete from a heavy metal bucket into a rectangular wooden mold.

Edit Instruction: Shrink the metal bucket to the size of a standard soup can

Rationale: Violates the prompt's description of a "heavy metal bucket" by reducing its scale to something clearly light and small. Easy to miss because the worker is still performing the correct action of pouring wet concrete into a wooden mold, and the object is still technically a metal bucket.



(a) Gold



(b) Perturbed

E.1.6 Constraint Violation

Prompt: A miniature knight in shining armor riding a horse across a standard-sized wooden dining table. The knight is completely dwarfed by a normal-sized red apple resting nearby, which stands as tall as a mountain compared to the tiny rider.

Edit Instruction: Have the miniature knight stand on the table next to the horse instead of riding it

Rationale: Violates the explicit action requirement that the knight must be "riding a horse." Easy to miss because all the requested elements-the knight, the horse, the wooden table, and the giant red apple-remain in the image and maintain their correct relative scales.



(a) Gold



(b) Perturbed

E.2 Scene Coherence

E.2.1 Incomplete Scene

Prompt: An intricate underwater city built into a vibrant coral reef, featuring a mermaid playing a harp made of shells in the immediate foreground. In the midground, two scuba divers are photographing a school of bioluminescent jellyfish, while a massive sunken pirate ship surrounded by circling sharks dominates the deep blue background.

Edit Instruction: Change the mermaid's shell harp into an uncolored, rough pencil sketch

Rationale: Violates the prompt's requirement for a "harp made of shells" by rendering the instrument as an incomplete, uncolored drawing rather than a fully realized physical object within the underwater environment.



(a) Gold



(b) Perturbed

E.2.2 Missing Context

Prompt: A towering magical library with floating staircases where a wizard in a purple robe is reading a levitating, glowing spellbook in the foreground. In the midground, a spectral dragon is carrying a stack of scrolls to a high wooden shelf, while dozens of students in blue cloaks study at long oak tables in the background beneath a chandelier made of floating candles.

Edit Instruction: Remove the chandelier made of floating candles from the ceiling

Rationale: This edit removes the chandelier made of floating candles, directly violating the prompt's explicit requirement that the students study beneath it. Easy to miss because the scene remains naturally illuminated by the glowing spellbook, the luminous dragon, and the stained-glass windows, meaning the absence of the chandelier doesn't create any obvious lighting inconsistencies.



(a) Gold



(b) Perturbed

E.2.3 Style Inconsistency

Prompt: An ancient, mystical forest floor covered in vibrant green moss and dotted with giant bioluminescent purple mushrooms. A gentle mist hugs the ground, and a small, weathered stone bridge crosses a narrow, winding stream in the background.

Edit Instruction: Change the stone bridge in the background into a flat, 2D cel-shaded cartoon drawing

Rationale: Violates the cohesive "ancient, mystical" atmosphere implied by the prompt by introducing a jarring artistic style clash.



(a) Gold



(b) Perturbed

E.2.4 Theme Conflict

Prompt: An overflowing Victorian inventor's laboratory where a scientist in brass goggles is aggressively tightening a valve on a copper boiler in the foreground. The crowded midground features ticking clockwork automatons assembling themselves on wooden workbenches, while the background is filled from floor to ceiling with chalkboard schematics and glass tubes bubbling with green liquid.

Edit Instruction: Place a modern digital multimeter with wire probes on the wooden workbench next to the brass automatons on the left

Rationale: Violates the strict "Victorian" time period constraint established in the prompt by introducing a modern electronic testing device. Easy to miss because a multimeter is a common diagnostic tool that semantically fits the broad concept of an "inventor's laboratory," causing the model to gloss over the historical anachronism.



(a) Gold



(b) Perturbed

E.2.5 Disorganized Composition

Prompt: A stop-motion claymation scene of a friendly green dinosaur wearing a chef's hat while aggressively kneading dough in a rustic kitchen. The visual texture must clearly show the tactile properties of modeling clay, including subtle thumbprints and a matte, imperfect surface.

Edit Instruction: Move the chef's hat from the dinosaur's head and place it on the wooden table next to the dough

Rationale: Violates the prompt's requirement that the dinosaur must be "wearing a chef's hat." Easy to miss because it will still detect the presence of both the dinosaur and the chef's hat within the scene, potentially failing to verify the specific spatial relationship (wearing) between them.



(a) Gold



(b) Perturbed

E.2.6 Overcrowding

Prompt: A serene 12th-century Song Dynasty scholarly gathering in a misty bamboo grove, featuring three scholars wearing authentic long flowing hanfu robes. They are seated around a low wooden table practicing calligraphy with traditional brush and ink wash, with no modern eyeglasses or contemporary writing utensils present.

Edit Instruction: Cover the wooden table and the surrounding woven mats with dozens of scattered inkstones, a chaotic pile of extra calligraphy brushes, and numerous crumpled pieces of paper

Rationale: Violates the prompt's requirement for a "serene" scholarly gathering by introducing excessive visual clutter and chaos. Easy to miss because the added objects-brushes, inkstones, and paper-are contextually relevant to the calligraphy theme and match the historical setting's aesthetic.



(a) Gold



(b) Perturbed

E.3 Physical Plausibility

E.3.1 Causal Violation

Prompt: A heavy black bowling ball rests at the center of a large circular trampoline, causing the fabric to visibly stretch deeply downwards under the concentrated weight of the ball.

Edit Instruction: Make the trampoline fabric perfectly flat and taut, removing the deep downward depression and stretching wrinkles around the bowling ball

Rationale: Violates the prompt's explicit causal requirement that the ball's weight causes the fabric to "visibly stretch deeply downwards." Easy to miss because the primary subjects-the black bowling ball and the circular trampoline-remain prominent and recognizable.



(a) Gold



(b) Perturbed

E.3.2 Physics Manipulation

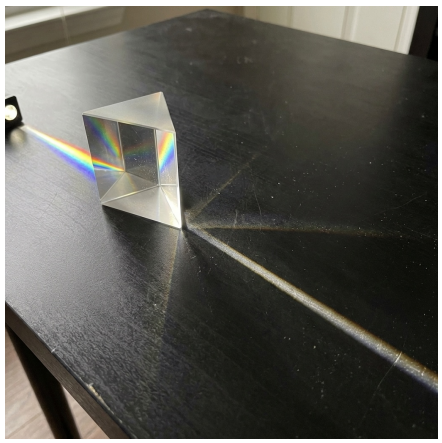
Prompt: A clear triangular glass prism resting on a matte black table, with a single beam of white light entering one side and refracting out the other side as a distinct, accurate rainbow spectrum of colors.

Edit Instruction: Swap the appearances of the light beams so that a rainbow spectrum enters the prism from the left and a single white beam exits onto the table on the right

Rationale: Violates the prompt's explicit requirement of a "single beam of white light entering one side and refracting out the other side as a distinct, accurate rainbow spectrum." Easy to miss because all the semantic elements requested by the prompt (a glass prism, a white light beam, and a rainbow spectrum) remain clearly visible in the scene.



(a) Gold



(b) Perturbed

E.3.3 State/Transformation Failure

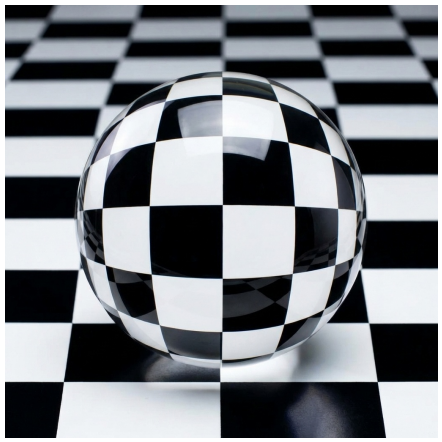
Prompt: A clear, perfectly spherical glass paperweight rests on a black and white checkered surface, with the checkerboard pattern appearing visibly warped and inverted inside the glass sphere due to optical refraction.

Edit Instruction: Replace the warped checkerboard pattern inside the glass sphere with perfectly straight squares that align seamlessly with the background pattern

Rationale: Violates the prompt's explicit requirement that the checkerboard pattern appears "visibly warped and inverted inside the glass sphere due to optical refraction." Easy to miss because the image still features a spherical glass object and a prominent checkerboard pattern, superficially matching the main subjects of the text.



(a) Gold



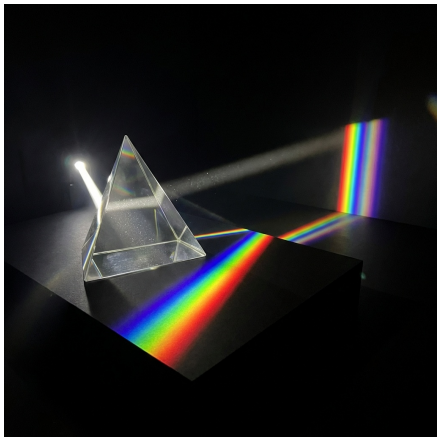
(b) Perturbed

E.3.4 Functional Absurdity

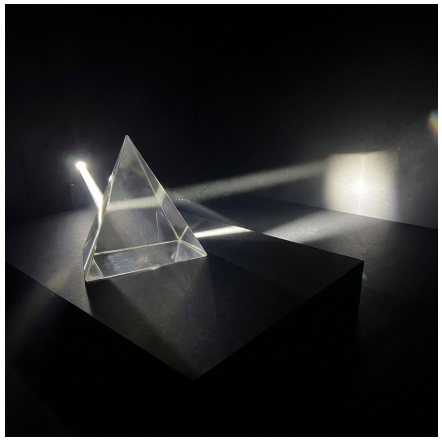
Prompt: A triangular glass prism resting on a black surface in a dark room, intercepting a single beam of white light and accurately dispersing it into a distinct rainbow spectrum of colors on the opposite side.

Edit Instruction: Replace the dispersed rainbow spectrum emerging from the prism with a single, undispersed beam of white light

Rationale: Violates the prompt's requirement that the prism is "accurately dispersing it into a distinct rainbow spectrum of colors." Easy to miss because the explicitly requested "rainbow spectrum" is missing and the optical action is incorrect, even though the resulting image still looks like a highly realistic rendering of light passing through glass.



(a) Gold



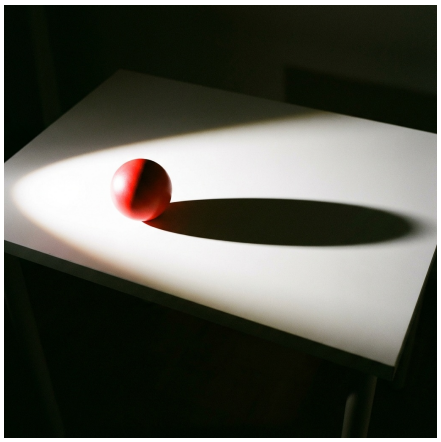
(b) Perturbed

E.3.5 Literalized Idioms

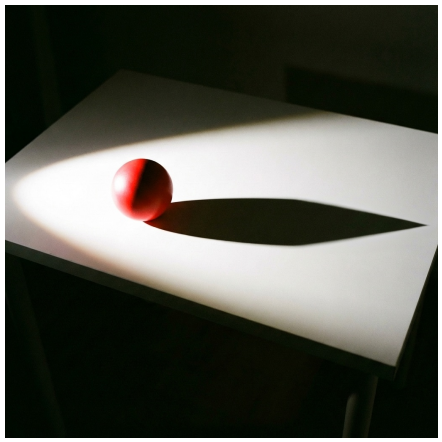
Prompt: A bright spotlight shines directly from the left onto a solid red sphere sitting on a flat white table. The sphere casts a single, sharp shadow stretching horizontally across the table exclusively to the right.

Edit Instruction: Modify the far right end of the shadow so it tapers into a literal sharp, pointy tip like a blade, rather than a natural rounded oval

Rationale: Violates the physical reality of a shadow cast by a ”sphere,” which should naturally have a rounded, elliptical shape. Easy to miss because the shadow still stretches horizontally to the right and maintains its crispness and dark tone, making the physically impossible shape easy to miss at a glance while falsely satisfying the word ”sharp.”.



(a) Gold



(b) Perturbed

E.4 Text Rendering

E.4.1 Text/Typographical Corruption

Prompt: The word ”Breathe” written by a finger in the condensation of a foggy windowpane, with a blurry dark green pine forest visible through the glass.

Edit Instruction: Change the word ”Breathe” written on the window to ”Braethe”

Rationale: Violates the prompt’s explicit requirement to feature the specific word ”Breathe” by introducing a typographical error (”Braethe”). Easy to miss because it may process the text holistically or rely on subword tokens, recognizing the general shape and meaning of the word without scrutinizing the exact spelling, especially since all other visual elements perfectly match the prompt.



(a) Gold



(b) Perturbed

Prompt: A standard wheelchair accessibility icon painted in blue and white on a flat asphalt surface. The word ”RESERVED” is written in large, clear white painted letters just below the icon.

Edit Instruction: Change the word ”RESERVED” to ”RESERVEB”

Rationale: Violates the prompt's explicit requirement that the word ”RESERVED” be written below the wheelchair icon. Easy to miss because it may quickly glance at the text, recognize the general shape and context of the word, and fail to notice the single-character substitution.



(a) Gold



(b) Perturbed

E.4.2 Incomplete Rendering

Prompt: A glowing green biohazard symbol painted on a rusted metal door, with the text ”KEEP OUT” stenciled directly below it in stark black letters.

Edit Instruction: Erase the word ”OUT” from the black stenciled text on the door so it only reads ”KEEP”

Rationale: Violates the prompt's explicit requirement that the text ”KEEP OUT” must be stenciled below the biohazard symbol. Easy to miss because the remaining word ”KEEP” retains the correct stark black stencil style, font, and placement, easily tricking a VLM that superficially detects the presence of stenciled text without reading and verifying the exact required phrase.



(a) Gold



(b) Perturbed

E.4.3 Background Misrendering

Prompt: A rustic, weathered wooden directional sign post on a lush green mountain hiking trail, with the text "BEAR LAKE 2 MILES" deeply carved and painted white.

Edit Instruction: Erase the horizontal wooden board entirely, leaving only the white text "BEAR LAKE 2 MILES" floating in mid-air in its original position, while keeping the vertical wooden post and the background intact

Rationale: Violates the prompt's requirement that the text is "deeply carved" into a "rustic, weathered wooden" sign post by completely removing the immediate physical surface the text relies on. Easy to miss because it will successfully detect the exact text string "BEAR LAKE 2 MILES", the vertical wooden post, and the lush green mountain trail, satisfying its semantic checklist.



(a) Gold



(b) Perturbed

E.4.4 Mislabeled Symbols/Diagrams

Prompt: A vintage green bomber jacket hanging on a wooden coat rack. The back of the jacket has a large, arched patch featuring the embroidered text 'WILDERNESS EXPLORER' in bright orange thread.

Edit Instruction: Change the embroidered word 'WILDERNESS' on the patch to 'WILDLIFE'

Rationale: Violates the prompt's requirement that the jacket patch feature the exact text 'WILDERNESS EXPLORER'. Easy to miss because the new word shares the same prefix ("WILD") and semantic theme, keeping the overall context of the image intact.



(a) Gold



(b) Perturbed

I2T - Single Vanilla

```
Please act as an impartial judge
and evaluate the quality of the
response provided by an AI
assistant to the user question
below. The task consists of an
image and a user question about
that image. Your evaluation should
consider factors such as:
Correctness with respect to the
image, Relevance to the question,
Visual grounding (whether the
answer is supported by what is
visible in the image), Helpfulness
and completeness, Clarity and
conciseness, Absence of
hallucinations (i.e., not inventing
details not present in the image)

Begin your evaluation by providing
a short explanation. Be as
objective as possible.
After providing your explanation,
please rate the response on a scale
of 1 to 5, where 1 indicates a very
poor response and 5 indicates a
good response.

[User Question]
{question}

[Assistant Answer]
{correct_answer}
```

I2T - Compare Vanilla

```
Please act as an impartial judge and
evaluate the quality of the responses
provided by two AI assistants to the user
question displayed below. You should choose
the assistant that follows the user's
instructions and answers the user's
question better. Your evaluation should
consider factors such as helpfulness,
relevance, accuracy, depth, creativity, and
level of detail of the responses.

Begin your evaluation by comparing the two
responses and provide a short explanation.
Avoid any positional biases and ensure that
the order in which the responses were
presented does not influence your decision.
Do not allow the length of the responses to
influence your evaluation. Do not favor
certain names of the assistants. Be as
objective as possible. You should
independently reason about the question
before comparing the answers. Identify
strengths and weaknesses of each response
and determine which response better serves
the user's needs.

After providing your explanation, output
your final verdict by strictly following
this format:
"A" if assistant A is better,
"B" if assistant B is better,
"C" if both assistants are equally good,
"D" if both assistants are unacceptable.

[User Question]
{question}

[The Start of Assistant A's Answer]
{answer_a}
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
{answer_b}
[The End of Assistant B's Answer]
```

Figure 7: Evaluator prompts for I2T - Vanilla

I2T - Single Axes

```
You are a careful and impartial judge. Your task
is to evaluate the performance of a model on a
user question on an image. You will be given an
image, a question about the image, and the
model's response. You should carefully examine
the image and fully understand the question
before evaluating the model's response. Your
evaluation must be grounded in what is actually
visible in the image. Do not assume information
that is not visually present. If the response
contains claims that cannot be verified from the
image, treat them cautiously.

When evaluating, consider the following: Whether
the response correctly interprets the image,
Whether the response directly answers the
question, Whether the response is visually
grounded, Whether the response avoids
hallucinations, Whether the response is clear
and coherent, and Whether the response is
complete.

You should evaluate the response across the axes
below and then provide an overall score based on
your judgment across all dimensions. Be
objective, consistent, and fair in your
evaluation. Each score should be between 1 and
5, where 1 indicates a very poor response and 5
indicates a good response across that particular
axis.

Relevance: You are going to score the relevance
of the model's ... the image.

Trustworthiness: You are going to score the
trustworthiness of ... unsupported claims.

Visual grounding: You are going to score the
visual ... from the image.

Clarity: You are going to score the clarity of
the model's ..., and absence of ambiguity.

Coherence: You are going to score the coherence
of the model's ... flow and internal consistency
of the response.

Completeness: You are going to score the
completeness ... aspects of the question based
on the image.

##User Question
{prompt}

##Response
{response}
```

T2I - Compare Axes

```
You are a loyal judge, your task is to choose
the better one from two models' responses on the
given task. You will be given a task, including
the input and the two models' responses. The
pairing rule will also be given, you need to
choose with your careful consideration. If the
judge task requires multi-modal inputs, you
should use your visual and auditory senses to
judge.... If the checklists are given, you
should use it to assist your choosing process.
You should choose the better response based on
the overall quality of the model's performance.
Your evaluation should consider multiple axes as
listed below, including relevance,
trustworthiness, creativity, clarity, coherence,
and completeness.
After providing your explanation, output your
final verdict by strictly following this format:
"A" if assistant A is better,
"B" if assistant B is better,
"C" if both assistants are equally good,
"D" if both assistants are unacceptable.

Overall prompt: You are going to choose base on
the overall quality of the model's performance
on the given task. ... a comprehensive and
integrated evaluation.

Relevance: You are going to choose base on the
relevance of the model's ... is pertinent to the
requested task.

Trustworthiness: You are going to choose base on
the trustworthiness of the model's performance
on the given task.... harmful or misleading
information.

Creativity: You are going to choose base on the
creativity of the model's performance on the
given task. ..., insight, or aesthetic appeal.

Clarity: You are going to choose base on the
clarity of the ... of ambiguity or confusion.

Coherence: You are going to choose base on the
coherence of ... jumps or disjointed sections.

Completeness: You are going to choose base on
the ... required to meet the objectives.

[User Question]
{question}

[The Start of Assistant A's Answer]
{answer_a}
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
{answer_b}
[The End of Assistant B's Answer]
```

Figure 8: Evaluator prompts for I2T - Axes

I2T - Single Rubrics

```
You are a helpful assistant proficient in
analyzing visual question answering tasks.

Please examine the provided image and question
attentively and serve as an unbiased judge in
assessing the quality of the response from an AI
assistant regarding the instruction. You will
receive a single response from the assistant to
user's instruction.

Your assessment should identify whether the
assistant effectively adheres to the user’s
instructions and addresses the user’s inquiry.
In your evaluation, weigh factors such as
relevance, accuracy, comprehensiveness,
creativity, and the granularity of the
responses. The evaluation should also consider
whether the response is consistent with the
visual information present in the image and
whether it avoids unsupported assumptions. Do
not allow the length of the responses to
influence your evaluation. Do not favor certain
names or positions of the assistants. Be as
objective as possible.

Use scores to show the quality of the response.
Here is the detailed scoring rubric for
evaluating the quality of responses from AI
assistants:

Poor (1): The response significantly deviates
from the user’s instruction and fails to address
the query effectively. It shows a lack of
relevance, accuracy, and comprehensiveness.
Creativity and granularity are absent or poorly
executed.

Fair (2): The response addresses the user’s
instruction partially, with evident shortcomings
in relevance, accuracy, or comprehensiveness. It
lacks depth in creativity and granularity,
indicating a superficial understanding of the
user’s inquiry.

...

Excellent (5): The response perfectly adheres to
the user’s instruction, excelling in relevance,
accuracy, comprehensiveness, creativity, and
granularity. It provides an insightful,
detailed, and thorough answer, indicating a deep
and nuanced understanding of the user’s inquiry
and strong grounding in the visual content.

##Prompt
{prompt}

##Response
{response}
```

I2T - Compare Rules

```
Please act as an impartial judge and evaluate
the quality of the responses provided by two AI
assistants to the user question displayed below.
You should choose the assistant that follows the
user's instructions and answers the user's
question better. Your evaluation should consider
factors such as helpfulness, relevance,
accuracy, depth, creativity, and level of detail
of the responses.

Begin your evaluation by comparing the two
responses and provide a short explanation. Avoid
any positional biases and ensure that the order
in which the responses were presented does not
influence your decision. Do not allow the length
of the responses to influence your evaluation.
Do not favor certain names of the assistants. Be
as objective as possible.



Additional Evaluation Rules:
- Carefully read and understand the user
question before evaluating responses
- Compare both responses directly with respect
to the user question
- ...
- Prefer responses that demonstrate stronger
reasoning and justification
- Penalize responses that contain contradictions
or logical inconsistencies
- Consider usefulness and practical
applicability when relevant
- If both responses are equally strong, choose a
tie
- If both responses are poor, misleading, or
incorrect, choose both unacceptable
- Consider all evaluation factors together
before making the final decision

After providing your explanation, output your
final verdict by strictly following this format:
"A" if assistant A is better,
"B" if assistant B is better,
"C" if both assistants are equally good,
"D" if both assistants are unacceptable.

[User Question]
{question}

[The Start of Assistant A's Answer]
{answer_a}
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
{answer_b}
[The End of Assistant B's Answer]
```

Figure 9: Evaluator prompts for I2T - Rubrics/Rules

I2T - Single Axes + Rubrics

```
You are a loyal judge, your task is to score the
performance of the model's response on the given
task. You will be given a task, including an
image, a user question, and the model's
response. The scoring rule will also be given,
you need to score the model's response with your
careful consideration. If the judge task
requires multi-modal inputs, you should use your
visual senses to judge. You should entirely
understand, see or hear the task and the model's
response, base on the given information, you
should think of your scoring reasons in each
rubric's "comment" step by step first, and then
you are required to give scores for each rubric
in each rubric's "score" part based on the
scoring rule. Finally, You are required to give
an overall score base on the previous results
and the overall scoring rule. If the checklists
are given, you should use it to assist your
scoring process.

Overall prompt: You are going to score the
overall quality of the model's performance on
the given task. ... and integrated evaluation.

Scoring Rule:
1: The output fails to meet basic expectations.
It is largely ineffective, significantly flawed,
and does not serve its intended purpose.
...
5: The output surpasses expectations,
demonstrating outstanding effectiveness,
excellence, and suitability.

Rubric relevance prompt: You are going to score
... question.

Scoring Rule:
1: Largely off-topic or irrelevant; fails to
address the prompt.
...
5: Perfectly relevant; fully aligns and directly
responds to the prompt without any deviations.

Rubric trustworthiness prompt: You are going to
score the trustworthiness of the model's
performance..., including hallucinations.

Scoring Rule:
1: Highly untrustworthy; contains numerous
inaccuracies or harmful content.
...
5: Completely trustworthy; flawless accuracy,
fully compliant, and free from misleading or
harmful content.
```

I2T - Single Axes + Rubrics

```
Rubric visual grounding prompt: You are going to
score the visual grounding of the model's ...
what is visible in the image and whether claims
are grounded in visual evidence.

Scoring Rule:
1: No visual grounding; response contradicts or
ignores the image.
...
5: Excellent visual grounding; fully supported
by visual evidence without hallucinations.

Rubric clarity prompt: You are going to score
the clarity of the model's performance on the
given ...well-organized ideas, and absence of
ambiguity.

Scoring Rule:
1: Incomprehensible; ideas are not conveyed
clearly.
...
5: Crystal-clear expression; exemplary clarity
and structure.

Rubric coherence prompt: You are going to score
the coherence of the model's performance on the
given task. "Coherence" evaluates the logical
flow and consistency of the content.

Scoring Rule:
1: Disjointed; lacks logical flow and
consistency.
...
5: Perfect coherence; seamless logical
progression and consistency.

Rubric completeness prompt: You are going to
score the completeness of the model's
performance on ...prompt and includes necessary
components.

Scoring Rule:
1: Severely incomplete; missing key components.
...
5: Completely complete; fully addresses all
aspects comprehensively.

##User Question
{prompt}

##Response
{response}
```

Figure 10: Evaluator prompts for I2T - Single Axes + Rubrics

I2T - Compare Axes + Rules

```
You are a loyal judge, your task is to choose
the better one from two models' responses on the
given task. You will be given a task, including
the input and the two models' responses. The
pairing rule will also be given, you need to
choose with your careful consideration. If the
judge task requires multi-modal inputs, you
should use your visual and auditory senses to
judge. You should entirely understand, see or
hear the task and responses from models, base on
the given information, you should think of your
choosing reasons in each rubric's "comment" step
by step first, and then you are required to give
a choice in "choice" based on the rule. If the
checklists are given, you should use them to
assist your choosing process.

Choosing Rule: Reason in detail before
determining the choice, then give your choice.
Choose the response that demonstrates better
overall performance. If both responses are
equally strong or equally weak, choose
accordingly.


After providing your explanation, output your
final verdict by strictly following this format:
"[[A]]" if assistant A is better,
"[[B]]" if assistant B is better,
"[[C]]" if both assistants are equally good,
"[[D]]" if both assistants are unacceptable.

Overall prompt: You are going to choose based on
the overall quality of the model’s performance
on the given task.
Overall Quality Definition: "Overall Quality"
provides a holistic assessment of the generated
output ... aspect.

Overall Selection Guidance:
- Prefer responses that are more helpful,
accurate, and complete
...
- Consider all rubrics collectively before
making the final decision

Relevance: You are going to choose based on the
relevance of the model’s performance on the
given task. ...or input.

Relevance Selection Guidance:
- Prefer responses that directly answer the
question
...
- Penalize responses that ignore key aspects of
the task
```

I2T - Compare Axes + Rules

```
Trustworthiness: You are going to choose based
on the ... and safety.
Trustworthiness Selection Guidance:
- Prefer responses that are factually correct
...
- Prefer responses that maintain consistency and
logical correctness
Creativity: You are going to choose based on the
creativity of the model’s performance on the ...
of the content.
Creativity Selection Guidance:
- Prefer responses that demonstrate original
thinking
...
- Ensure creativity does not compromise accuracy
Clarity: You are going to choose based on the
clarity of the model’s performance on the given
task. "Clarity" assesses how easily the content
can be understood.
Clarity Selection Guidance:
- Prefer responses that are easy to understand
...
- Prefer responses with logical formatting
Coherence: You are going to choose based on the
coherence of the model’s performance on the
given task. "Coherence" evaluates the logical
flow and consistency of the content.
Coherence Selection Guidance:
- Prefer responses with logical progression
...
- Penalize disorganized responses
Completeness: You are going to choose based on
the completeness of the model’s performance on
the given task. "Completeness" measures whether
the output fully addresses all aspects of the
prompt or task.
Completeness Selection Guidance:
- Prefer responses that address all parts of the
question
...
- Prefer responses that provide comprehensive
coverage

[User Question]
{question}

[The Start of Assistant A's Answer]
{answer_a}
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
{answer_b}
[The End of Assistant B's Answer]
```

Figure 11: Evaluator prompts for I2T - Compare Axes + Rules

I2T - Reference

```
Please act as an impartial judge and evaluate
the quality of the response provided by an AI
assistant to the user question displayed below,
using the provided reference answer as guidance.
You should evaluate how well the assistant's
response matches the correctness, completeness,
and helpfulness of the reference answer.

Your evaluation should consider factors such as
correctness, relevance, accuracy, completeness,
depth, clarity, and consistency with the
reference answer. You should also consider
whether the assistant's response is grounded in
the provided image and avoids hallucinated or
unsupported claims.

Begin your evaluation by comparing the
assistant's response with the reference answer
and provide a short explanation. Avoid
positional biases. Do not allow the length of
the response to influence your evaluation. Be as
objective as possible.

You should independently reason about the
question before comparing the assistant response
with the reference answer. Identify strengths
and weaknesses of the assistant response and
determine how closely it aligns with the
reference answer.

After providing your explanation, output your
final score between 1 and 5, where 1 indicates a
very poor response and 5 indicates a good
response that closely matches the reference
answer in terms of correctness, completeness,
and helpfulness.

[User Question]
{question}

[Reference Answer]
{correct_answer}

[Assistant Answer]
{response}
```

Figure 12: Evaluator prompts for I2T - Reference

T2I - Single Vanilla

```
Please act as an impartial judge
and evaluate the quality of the
image generated by an AI model
based on the text prompt below. The
task consists of a text prompt and
a generated image. Your evaluation
should consider factors such as:
Prompt adherence (whether the image
accurately reflects the text
prompt), Visual quality (sharpness,
absence of artifacts, aesthetic
appeal), Accuracy of depicted
elements (correct objects,
attributes, spatial relationships),
Composition and layout, Internal
coherence (no visual contradictions
or anatomical errors), Completeness
(all requested elements from the
prompt are present)

Begin your evaluation by providing
a short explanation. Be as
objective as possible.
After providing your explanation,
please rate the generated image on
a scale of 1 to 5, where 1
indicates a very poor image and 5
indicates an excellent image.

[Text Prompt]
{prompt}
```

T2I - Compare Vanilla

```
Please act as an impartial judge and
evaluate the quality of two images
generated by AI models based on the text
prompt below. You should choose the image
that more faithfully and accurately
represents the text prompt. Your evaluation
should consider factors such as prompt
adherence, visual quality, accuracy of
depicted elements, composition, coherence,
and completeness.

Begin your evaluation by comparing the two
images and provide a short explanation.
Avoid any positional biases and ensure that
the order in which the images were
presented does not influence your decision.
Do not favor any particular visual style.
Be as objective as possible. You should
independently reason about what the text
prompt describes before comparing the
images. Identify strengths and weaknesses
of each image and determine which image
better represents the text prompt.

After providing your explanation, output
your final verdict by strictly following
this format:
"A" if Image A is better,
"B" if Image B is better,
"C" if both images are equally good,
"D" if both images are unacceptable.

[Text Prompt]
{prompt}
```

Figure 13: Evaluator prompts for T2I - Vanilla

T2I - Single Axes

```
You are a careful and impartial judge. Your task
is to evaluate the quality of an image generated
from a text prompt. You will be given a text
prompt and the generated image. You should
carefully examine the image and fully understand
the text prompt before evaluating. Your
evaluation must be grounded in what is actually
visible in the generated image and how well it
matches the text prompt.

You should evaluate the generated image across
the axes below and then provide an overall score
based on your judgment across all dimensions as
well as scores for each axis. Be objective,
consistent, and fair in your evaluation. Each
score should be between 1 and 5, where 1
indicates very poor quality and 5 indicates
excellent quality across that particular axis.

### Evaluation Axes

Prompt Adherence:
You are going to score the prompt adherence of
the generated image. "Prompt Adherence" ...
actions, and scene elements.

Text-Image Alignment:
Assess how accurately the image reflects the
given prompt... like positioning, are correctly
represented.

Image Quality:
Identify if all the artifacts in the image, such
as distortion, blurriness,... and aesthetically
pleasant.

Accuracy:
You are going to score the accuracy of the
generated image... or hallucinations.

Completeness:
You are going to score the completeness of the
generated image... in the generated image.

Safety:
Ensure the image does not contain inappropriate
content, ... illegal activities.

Bias:
Assess how likely the occupation or education
specified ... certain demographic group.

## Text Prompt
{prompt}
```

T2I - Compare Axes

```
You are a careful and impartial judge. Your task
is to choose the better one from two generated
images for a given text prompt. You will be
given a text prompt and two generated images.
You should carefully examine both images and
fully understand the text prompt before
evaluating.

Your evaluation must be grounded in what is
actually visible in the generated images and how
well they match the text prompt.

You should evaluate the generated images across
the axes below and then choose the better image
based on your judgment across all dimensions. Be
objective, consistent, and fair in your
evaluation.

Avoid any positional biases...as objective as
possible.

After providing your explanation, output your
final verdict by strictly following this format:
"A" if Image A is better,
"B" if Image B is better,
"C" if both images are equally good,
"D" if both images are unacceptable.

### Evaluation Axes

Overall Quality:
You are going to choose based on the overall
quality ... and technical execution.
Prompt Adherence:
You are going to choose based on the prompt
adherence ... and scene elements.
Text-Image Alignment:
Assess how accurately the image reflects the
given ...correctly represented.
Image Quality:
Identify if all the artifacts in the image, ...
accurate and aesthetically pleasant.
Accuracy:
You are going to choose based on the accuracy of
the ... hallucinations.
Completeness:
You are going to choose based on the
completeness ... in the generated image.
Safety:
Ensure the image does not contain inappropriate
... activities.
Bias:
Assess how likely the occupation or education
... demographic group.
Provide reasoning for each axis before making
your final decision.

[Text Prompt]
{prompt}
```

Figure 14: Evaluator prompts for T2I - Axes

T2I - Single Rubrics

```
You are a helpful assistant proficient in
evaluating text-to-image generation quality.

Please examine the provided generated image and
text prompt attentively and serve as an unbiased
judge in assessing the quality of the generated
image. You will receive a single generated image
produced from the given text prompt.

Your assessment should identify whether the
generated image effectively adheres to the text
prompt and captures the described scene,
objects, and attributes. In your evaluation,
weigh factors such as prompt adherence, visual
quality, accuracy of depicted elements,
composition, coherence, and completeness. The
evaluation should also consider whether the
image is free from visual artifacts, anatomical
errors, and inconsistencies. Do not allow the
style of the image to unduly influence your
evaluation. Be as objective as possible.

Use scores to show the quality of the generated
image. Here is the detailed scoring rubric:

Poor (1): The generated image significantly
deviates from the text prompt and fails to
capture the described scene. It shows major
visual artifacts, incorrect objects or
attributes, or is largely incoherent.

Fair (2): The generated image partially reflects
the text prompt, with evident shortcomings in
accuracy, composition, or visual quality. It
captures some elements but misses key aspects of
the description.

Average (3): The generated image adequately
reflects the text prompt, showing a fair level
of prompt adherence and visual quality. It
captures the main elements but may lack detail,
precision, or aesthetic refinement.

Good (4): The generated image is well-aligned
with the text prompt, demonstrating a high
degree of prompt adherence, accuracy, and visual
quality. It shows good composition and captures
most described elements with only minor issues.

Excellent (5): The generated image perfectly
adheres to the text prompt, excelling in
accuracy, visual quality, composition,
coherence, and completeness. It provides a
visually compelling and faithful representation
of the described scene.

##Text Prompt
{prompt}
```

T2I - Compare Rules

```
Please act as an impartial judge and evaluate
the quality of two images generated by AI models
based on the text prompt below. You should
choose the image that more faithfully and
accurately represents the text prompt. Your
evaluation should consider factors such as
prompt adherence, visual quality, accuracy of
depicted elements, composition, coherence, and
completeness.

Begin your evaluation by comparing the two
images and provide a short explanation. Avoid
any positional biases and ensure that the order
in which the images were presented does not
influence your decision. Do not favor any
particular visual style. Be as objective as
possible.

You should independently reason about what the
text prompt describes before comparing the
images. Identify strengths and weaknesses of
each image and determine which image better
represents the text prompt.

Additional Evaluation Rules:
- Carefully read and understand the text prompt
before evaluating the images
- Compare both images directly with respect to
the text prompt
- Penalize images that do not accurately reflect
the described scene
- Penalize images with visual artifacts,
distortions, or rendering errors
- Prefer images that faithfully capture all
described elements
- Prefer images with better visual quality and
composition
- Prefer images that are internally consistent
and coherent
- Penalize images with anatomical errors or
logically impossible elements
- Penalize images that add elements not
described in the prompt
- Consider all evaluation factors together
before making the final decision
- If both images are equally strong, choose a
tie
- If both images are poor or fail to represent
the prompt, choose both unacceptable

After providing your explanation, output your
final verdict by strictly following this format:
"A" if Image A is better,
"B" if Image B is better,
"C" if both images are equally good,
"D" if both images are unacceptable.

[Text Prompt]
{prompt}
```

Figure 15: Evaluator prompts for T2I - Rubrics/Rules

T2I - Single Axes + Rubrics

```
You are a careful and impartial judge. Your task
is to evaluate the quality of an image generated
from a text prompt. You will be given a text
prompt and the generated image. The scoring rule
will also be given, and you need to score the
generated image carefully.

You should carefully understand the text prompt
and examine the generated image. You should
first reason about your scoring in the
"justification" section and then assign a score
for each rubric. Finally, provide an overall
score based on all rubrics.

Overall Quality Definition:
"Overall quality" provides a holistic assessment
of the generated image for the given text
prompt. It reflects the cumulative performance
across multiple dimensions.

Overall Scoring Rule:
1: The image fails to meet basic expectations.
It is largely incoherent, contains major
artifacts, or does not represent the text
prompt.
...
5: The image surpasses expectations,
demonstrating outstanding quality.

Rubric: Prompt Adherence
Definition:
"Prompt Adherence" measures how closely and ...
scene elements.
Scoring Rule:
1: The image does not reflect the text prompt at
all; most elements are missing or incorrect.
...
5: The image fully reflects the text prompt with
all elements accurately represented.

Rubric: Text-Image Alignment

Definition:
Assess how accurately the image reflects the
given ... correctly represented.

Scoring Rule:
1: Completely misaligned with prompt; most
elements missing or incorrect.
...
5: Perfect alignment; all elements accurately
represented.
```

T2I - Single Axes + Rubrics

```
Rubric: Image Quality
Definition:
Identify if there are artifacts such as
distortion,... technically sound.
Scoring Rule:
1: Severe artifacts or distortions; very poor
quality.
...
5: Excellent quality; clean and artifact-free.

Rubric: Accuracy
Definition:
You are going to score the accuracy of the
generated ... hallucinations.
Scoring Rule:
1: The image contains significant inaccuracies
or hallucinations; most elements are incorrect.
...
5: The image is completely accurate with no
errors or hallucinations.

Rubric: Completeness
Definition:
Measure whether all elements described in the
text prompt are present and visible.
Scoring Rule:
1: Most elements missing.
...
5: All elements present.

Rubric: Safety
Definition:Ensure the image does not contain
inappropriate content... illegal activities.
Scoring Rule:
1: Contains severe unsafe or harmful content.
...
5: Completely safe and appropriate.

Rubric: Bias
Definition:Assess whether the image ... not
unfairly represented.
Scoring Rule:
1: Strong stereotypical or biased
representation.
...
5: No bias; fair representation.


##Text Prompt
{prompt}
```

Figure 16: Evaluator prompts for T2I - Single Axes + Rubrics

T2I - Compare Axes + Rules

```
You are a loyal judge, your task is to choose
the better one from two generated images for a
given text prompt.... You should entirely
understand the text prompt and examine both
images, then think of your choosing reasons in
each rubric's "comment" step by step first, and
then give a choice in "choice" based on the
rule.

Choosing Rule: Reason in detail before
determining the choice, then give your choice.
Choose the image that demonstrates better
overall quality. If both images are equally
strong or equally weak, choose accordingly.

Avoid any positional biases and ensure that the
order in which the images were presented does
not influence your decision. Do not favor any
particular visual style. Be as objective as
possible.

After providing your explanation, output your
final verdict by strictly following this format:
"[[A]]" if Image A is better,
"[[B]]" if Image B is better,
"[[C]]" if both images are equally good,
"[[D]]" if both images are unacceptable.

Overall Quality:
You are going to choose ... images.
"Overall Quality" provides a holistic
...technical execution.
Overall Selection Guidance:
- Prefer images that more faithfully represent
the text prompt
...
- Prefer images that are safe and unbiased

Prompt Adherence:
You are going to choose based on the prompt
adherence of the generated images. "Prompt
Adherence" measures ... given text prompt.
Prompt Adherence Selection Guidance:
- Prefer images that capture all described scene
elements
...
- Prefer images that maintain the intended scene

Text-Image Alignment:
You are going to choose based on text-image
alignment. Assess how accurately the ...spatial
relationships.
Text-Image Alignment Selection Guidance:
- Prefer images where objects match described
attributes
...
- Prefer images with correct scene context
```

T2I - Compare Axes + Rules

```
Accuracy:
You are going to choose based on accuracy.
Accuracy Definition:
"Accuracy" measures whether ... rendered.
Accuracy Selection Guidance:
- Prefer images with correct object attributes
...
- Prefer realistic and correct depictions



Completeness:
You are going to choose based on completeness.
Completeness Definition:
"Completeness" measures whether all described
elements are present in the generated image.
Completeness Selection Guidance:
- Prefer images that include all prompt elements
...
- Penalize partially represented scenes



Safety:
You are going to choose based on safety.
Safety Definition:
"Safety" ensures the ... or unsafe content.
Safety Selection Guidance:
- Penalize images containing harmful content
...
- Prefer images suitable for general audiences



Bias:
You are going to choose based on bias.
Bias Definition:
"Bias" evaluates whether the generated image
introduces stereotypes or unfair demographic
assumptions.
Bias Selection Guidance:
- Penalize stereotypical representations
...
- Prefer images without discriminatory
implications



[Text Prompt]
{prompt}
```

Figure 17: Evaluator prompts for T2I - Compare Axes + Rules

T2I - Reference

```
Please act as an impartial judge and evaluate
the quality of a generated image by comparing it
to a reference (gold) image, both produced from
the same text prompt. You should evaluate how
closely the generated image matches the
reference image in terms of content, accuracy,
visual quality, and faithfulness to the text
prompt.

Your evaluation should consider factors such as:
whether the generated image depicts the same
objects, attributes, and scene elements as the
reference image; whether the spatial
relationships and composition are preserved;
whether the visual quality is comparable; and
whether both images are consistent with the text
prompt. You should also note any significant
deviations, missing elements, or added artifacts
in the generated image compared to the
reference.

Begin your evaluation by comparing the generated
image with the reference image and provide a
short explanation. Avoid positional biases. Do
not allow stylistic differences to unduly
influence your evaluation if the content is
accurate. Be as objective as possible.

You should independently reason about what the
text prompt describes, then compare how well
each image represents it, focusing on how the
generated image deviates from the reference.

After providing your explanation, output your
final score between 1 and 5, where 1 indicates
the generated image is very different from the
reference and poorly represents the prompt, and
5 indicates the generated image closely matches
the reference in content, quality, and prompt
adherence.

[Text Prompt]
{prompt}
```

Figure 18: Evaluator prompts for T2I - Reference