SKILL-ALIGNED ANNOTATION FOR RELIABLE TEXT-TO-IMAGE EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image (T2I) generation has advanced rapidly, making reliable evaluation critical as performance differences between models narrow. Existing evaluation practices typically apply uniform annotation mechanisms, such as Likert-scale or binary question answering (BQA), across heterogeneous evaluation skills, despite fundamental differences in their nature. In this work, we revisit T2I evaluation through the lens of *skill-aligned* annotation, where annotation strategies reflect the underlying characteristics of each evaluation skill. We systematically compare skill-aligned annotation against uniform baselines and show that it produces more consistent evaluation signals, with higher inter-annotator agreement and improved stability across models. In controlled comparisons with the same annotators, skill-aligned protocols raise inter-annotator agreement (Krippendorff's α) from as low as 0.39 to 0.82, cut annotator abstention from 63% to 2.5%, and stabilize with four to five annotators. Finally, we present an automated pipeline that instantiates the proposed evaluation protocol, enabling scalable and fine-grained evaluation with spatially grounded feedback. Our work highlights that improving the foundations of image evaluation can increase reliability and efficiency without simply scaling annotation effort. We hope this motivates further research on refining evaluation protocols as a central component of reliable model assessment.

1 INTRODUCTION

Text-to-Image (T2I) generation has recently undergone rapid progress with the emergence of diffusion models (Ho et al., 2020; Saharia et al., 2022; Podell et al., 2024; Esser et al., 2024). As performance differences between models continue to narrow, *reliable evaluation* becomes increasingly critical for guiding meaningful progress and identifying failure modes. A large body of work has therefore focused on evaluating T2I systems across multiple axes, including semantic fidelity, compositional correctness, and visual quality (Kirstain et al., 2023; Xu et al., 2023; Hu et al., 2023; Cho et al., 2024; Wiles et al., 2025). In practice, the gold standard for T2I evaluation relies on human feedback, where human annotators assess different evaluation skills (*e.g.* spatial arrangement, text rendering, visual quality) using annotation mechanisms such as Likert-scale ratings, ranking, or binary question answering (BQA). These mechanisms are typically applied *uniformly* across heterogeneous skills, despite fundamental differences in their nature, which raises a central question: "*Are standard annotation mechanisms uniformly effective in capturing heterogeneous skills in T2I generation?*"

Each of these uniform mechanisms carries implicit assumptions about what annotators know and what form of evidence a skill provides, and these assumptions break for many skills. For example, assessing landmark or style recognition via BQA presumes annotators possess sufficient prior knowledge, which is rarely verified. Conversely, compressing spatially localized artifacts into a single Likert score discards fine-grained evidence and can lead to disproportionate penalties for minor defects. These mismatches suggest that *annotation design is a critical but underexplored component of evaluation methodology*.

In this work, we revisit T2I evaluation from the perspective of *skill-aligned* annotation mechanisms, where these mechanisms are chosen to match the nature of each evaluation skill. To achieve this, we build an exhaustive skill taxonomy, starting from prior work, that covers both semantics and aesthetics, use it to systematically tag text prompts, and design annotation strategies tailored to each skill type. Crucially, we treat annotation design as a *methodological variable* and conduct controlled

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



Figure 1: Overview of the proposed Text-to-Image evaluation protocol using *skill-aligned* annotation, yielding more consistent and stable evaluation signals than uniform baselines.

comparisons between skill-aligned and uniform annotation protocols. To our knowledge, this is the first work to choose the annotation mechanism for each skill based on its measured reliability, rather than fixing one mechanism in advance. We show that skill-aligned annotation yields higher inter-annotator agreement and outcomes that stabilize with fewer annotators. Finally, to support large-scale evaluation, we instantiate the proposed protocol in an automated pipeline that enables scalable, fine-grained, and spatially grounded assessment, achieving strong alignment with human judgments across most skills.

Contributions (1) A controlled, within-subject comparison of skill-aligned protocols against the uniform mechanisms of current practice, which serve as our baselines: BQA as in TIFA, DSG, GenEval, and Gecko (Hu et al., 2023; Cho et al., 2024; Ghosh et al., 2023; Wiles et al., 2025), and global Likert scoring as in ImageReward and multi-dimensional preference models (Xu et al., 2023; Zhang et al., 2024). (2) Skill-aligned annotation protocols, namely brush-based, word-level with gap tokens, and anchor-based annotation, each selected per skill from measured reliability rather than fixed a priori. (3) An automated pipeline that instantiates the suite with VLM agents, and its per-skill agreement with human judgments.

We hope this work encourages a shift toward more principled evaluation design in T2I generation, emphasizing the role of annotation protocols in determining evaluation reliability.

2 Related Work

2.1 Evaluation Axes of T2I Generation

Early and widely adopted approaches formulate T2I evaluation as a global human preference task, where annotators rank images by overall appeal (Kirstain et al., 2023; Wu et al., 2023b;a; Ma et al., 2025); such signals capture aggregate preference but do not disentangle the factors behind it. A large body of work therefore targets *semantic fidelity*, decomposing a prompt into verifiable constraints such as object presence, counts, attributes, and spatial relations (Cho et al., 2024; Lin et al., 2024; Hu et al., 2023; Ghosh et al., 2023; Wiles et al., 2025; Li et al., 2024; Huang et al., 2025; Lu et al., 2023; Ku et al., 2024; Zhang et al., 2025), while complementary work covers perceptual axes such as realism, aesthetics, and visual artifacts, independently or jointly with semantic fidelity (Zhang et al., 2024; Otani et al., 2023; Hua et al., 2025; Xu et al., 2023; 2024; Lee et al., 2023). These works define *what* to evaluate and measure every axis with a single annotation mechanism chosen a priori. We instead take the evaluation axes as given and study *how* each one should be measured, selecting the annotation mechanism per skill from its measured reliability.

2.2 Annotation Strategies in T2I Evaluation

Human evaluation remains the gold standard for T2I generation. Comparative evaluation, where annotators choose between images from different models (Kirstain et al., 2023; Wu et al., 2023b;a;

Ma et al., 2025), reduces calibration requirements but depends on the chosen baselines, scales quadratically with the number of models, and does not reveal which aspect of an image drives a preference (Otani et al., 2023). We therefore focus on *absolute* evaluation, where individual images are assessed against predefined criteria with Likert ratings or binary question answering (BQA) (Cho et al., 2024; Hu et al., 2023; Ghosh et al., 2023; Xu et al., 2023; Wiles et al., 2025; Saxon et al., 2024), which supports skill-level attribution of errors but depends strongly on how the annotation mechanism is designed. Gecko (Wiles et al., 2025) shows that the annotation template affects agreement and model rankings for prompt–image alignment, but it applies each template uniformly to every prompt and thus asks which template is best overall rather than which suits each skill. RichHF (Liang et al., 2024) collects different feedback types for different aspects (point marks on implausible or misaligned regions, marked keywords for misalignment, and Likert scores for plausibility, alignment, and aesthetics) and trains a model to predict them; however, this scheme is fixed a priori and never compared against alternative mechanisms for the same skill. In contrast, we compare mechanisms head-to-head for each skill on identical images and annotators, introduce mechanisms for skills that alignment templates do not cover (artifacts, text rendering, and reference-dependent skills), and select the protocol per skill from measured agreement.

2.3 Automated Evaluation of T2I Models

To reduce the cost of human studies, prior work fine-tunes CLIP-based or related models as preference and scoring functions (Kirstain et al., 2023; Xu et al., 2023; Wu et al., 2023a; Zhang et al., 2024; Liang et al., 2024), or uses VQA models to answer semantic probe questions (Cho et al., 2024; Hu et al., 2023; Li et al., 2024; Lu et al., 2023); more recently, large Vision–Language Models (VLMs) serve as general-purpose judges (Xu et al., 2024; Ma et al., 2025; Hua et al., 2025; Hayes et al., 2025; Tu et al., 2025). These evaluators inherit the annotation protocol of the human data they are trained or validated on, often conflate axes, remain sensitive to prompt formulation, and typically apply one mechanism to every axis; MMIG-Bench (Hua et al., 2025) is a partial exception, using different models for low- and high-level properties. Our automated pipeline instead mirrors the skill-aligned protocols, assigning a dedicated mechanism or model to each skill, and we report its per-skill agreement with human judgments.

3 Evaluation Setup for Text-to-Image Generation

3.1 Skill Taxonomy for T2I Evaluation

Several prior works propose taxonomies for evaluating text-to-image (T2I) generation (Cho et al., 2024; Hu et al., 2023; Wiles et al., 2025; Lin et al., 2024). Building on these efforts, we extend the Gecko taxonomy (Wiles et al., 2025) to capture the key skills required for high-quality image generation, spanning both *semantic fidelity* and *aesthetic quality*. Starting from the Gecko skills, we added the perceptual skills that semantics-only taxonomies omit (visual artifacts and aesthetic quality), and then iterated between tagging the Gecko prompts, reviewing phrases left untagged or mis-assigned, and adding the missing skills until every prompt was covered. This loop surfaced seven semantic skills absent from Gecko (lighting, weather, viewpoint, camera, mood, time, and general scenes) and regrouped Gecko's skills into entities, attributes, and relative comparisons, adding sub-skills such as uncountable entities and tone and distance comparisons. Table 1 summarizes the taxonomy, and operational definitions and per-skill examples are provided in Appendix A. For evaluation prompts, we adopt the Gecko prompt set (Wiles et al., 2025) due to its broad coverage across existing benchmarks. In contrast to Gecko, which assigns a single skill–subskill pair per prompt, we annotate each prompt with all applicable skill and sub-skill combinations, which reduces the number of prompts required to cover the taxonomy.

3.2 Skill Tagging

Prior work explored tagging prompts using large language models (LLMs) (Hu et al., 2023; Cho et al., 2024; Li et al., 2024), followed by a second LLM that generates (BQA) probes to validate skill fulfillment. We adopt a similar approach but jointly perform tagging and question generation within a single LLM prompt (GPT-5). For each input prompt, the LLM is provided with the full skills taxonomy and asked to: (1) identify all relevant skill–subskill pairs, and (2) generate validation

Evaluation Skill	Sub-Skills	Uniform Baseline	Skill-Aligned (Ours)
Entities	Singular / Count / Uncountable		
Attributes	Color / Texture / Material / Shape / Scale		
Action	Standard / Unusual / Pose		
Arrangement	–		
Comparison	Scale / Tone / Distance / Count / Other		
Lighting	–	BQA	BQA
Weather	–		
View	–		
Camera	–		
Mood / Feeling	–		
Language Complexity	Negation / Color Stroop		
Time	Time of Day / Season / Year / Era		
Environment / Scene	Landmark / General		
Style	Artistic Style / Visual Medium	BQA	Anchor Likert
Named Entities	Character / Vehicle / Product / Artwork		
Text Rendering	Accuracy / Style / Numerical / Position	BQA / Likert	Word-Level
Artifacts	–	Likert	Brush
Aesthetic Quality	–	Likert	Likert

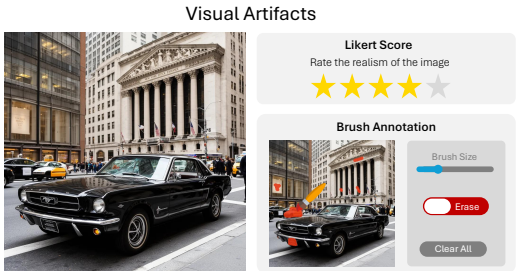
Table 1: Skill taxonomy with the uniform baseline mechanism of prior work and the skill-aligned protocol selected for each skill (Brush: Section 4.1; Word-Level: Section 4.2; Anchor Likert: Section 4.3).

questions. Following DSG (Cho et al., 2024), we structure the questions as dependency nodes. For example, for "A red car", the color question depends on the presence of the car. The LLM returns structured output that lists, for each skill–subskill pair, the prompt phrase, the validation question, its node type (presence, property, or relation), and the parent questions it depends on; the schema and further details are given in Appendix B. All tags and validation questions used in our experiments were manually verified before annotation, so tagger accuracy only matters when the pipeline is scaled without review. An overview of the tagging pipeline is shown in Figure 7.

4 Skill-Aligned Annotation

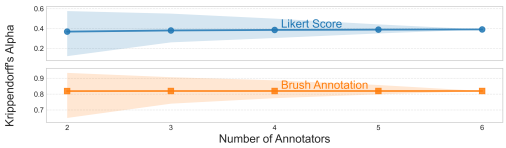
Given the skill taxonomy introduced in Section 3.1, we study how annotation protocols interact with different T2I skills. Our goal is to test whether protocols aligned with each skill provide more reliable evaluation signals than uniform annotation strategies. The baselines are the uniform mechanisms of current practice: BQA as in TIFA, DSG, GenEval, and Gecko (Hu et al., 2023; Cho et al., 2024; Ghosh et al., 2023; Wiles et al., 2025), and global Likert scoring as in ImageReward (Xu et al., 2023). Each skill-aligned protocol is compared with them head-to-head on identical prompts, images, and annotators.

As summarized in Table 1, *factual skills* such as object presence, counting, spatial arrangement, and temporal attributes are evaluated with binary question answering (BQA), since their evidence is a checkable fact that a yes/no answer captures adequately. For skills with less discrete evidence, such as visual artifacts, text rendering, landmarks, and artistic style, we compare multiple candidate annotation strategies. A fixed panel of six annotators, balanced by gender and from diverse cultural backgrounds, rates every image under every protocol, so protocol differences are not confounded with annotator differences. This fully crossed design gives each of the 90 images per skill (30 prompts, three models) six ratings under every protocol. The same panel produced both the lowest and the highest agreement values (*e.g.* $\alpha = 0.39$ vs. 0.82 on identical artifact images), so the gaps between protocols cannot be attributed to the panel. The gains come from what each protocol asks annotators to report rather than from who they are: brushing records the perceived defect instead of compressing it into a score, word tiles replace one holistic verdict with atomic checks, and anchors supply missing reference knowledge. Instructions were kept minimal (Appendix D), and agreement under skill-aligned protocols stabilizes with four to five annotators. For model diversity, we generate images



(a)

	Inter-Annotator Agreement (Krippendorff's α) ↑	EDR ($\geq 40\%$) ↓
Likert Score	0.39 [0.27, 0.50]	62.2%
Brush (Ours)	**0.82** [0.56, 0.89]	**0.0%**



(b)

Figure 2: Likert scoring (uniform baseline) vs. brush-based annotation (ours) for visual artifacts: inter-annotator agreement (Krippendorff's α, 95% bootstrap CI), extreme disagreement rate, and convergence.

using three open-source T2I models: FLUX.1-dev (Black Forest Labs, 2024), Z-Image-Turbo (Cai et al., 2025), and FLUX.2-dev (Black Forest Labs, 2025).

4.1 Visual Artifacts

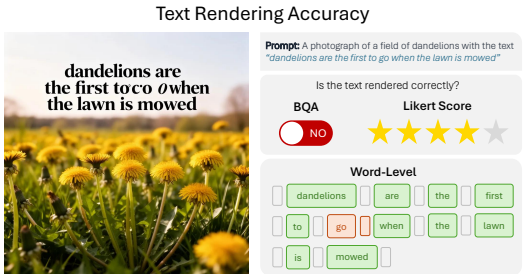
Visual artifacts are spatially localized but are often evaluated using global Likert ratings for realism or artifact severity (Xu et al., 2023; Zhang et al., 2024), which we take as the uniform baseline. To evaluate the reliability of this approach, we sample 30 prompts that commonly induce artifacts (humans, vehicles, animals, actions, and text rendering), generate images with the three models above, and ask annotators to rate each image on a 1–5 Likert scale reflecting perceived realism and absence of artifacts.

Figure 2b shows that Likert scoring yields low agreement, with Krippendorff's $\alpha = 0.39$ and 95% bootstrap confidence intervals (CI) of $[0.27, 0.50]$. All CIs are obtained by resampling images with replacement, so they capture the uncertainty due to the finite set of rated items. To further characterize disagreement, we define the **Extreme Disagreement Rate (EDR)** as the proportion of images for which the maximum annotator score difference exceeds a threshold τ. Let D_i denote the difference between the maximum and minimum score assigned to image i; we compute EDR $= \frac{1}{|V|}\sum_{i \in V} \mathbb{I}(D_i \geq \tau)$, where V is the set of images annotated by at least two annotators and $\mathbb{I}(\cdot)$ is the indicator function. We set τ to 40% of the scoring range, corresponding to two points on a 1–5 Likert scale: the smallest gap beyond ordinary calibration noise (a one-point threshold flags 98.9% of images); our conclusions hold at 20%–60% of the range (Appendix E). Under Likert scoring, the EDR reaches 62.2%, indicating substantial disagreement among annotators.

Skill-Aligned Protocol: Brush Annotation Because artifacts are spatially localized, we compare Likert scoring with region-based brush annotation, which is commonly used in artifact localization (Eldesokey et al., 2025; Kang et al., 2025; Zhang et al., 2023). Instead of assigning a global score, annotators mark image regions they perceive as artifacts. Unlike the point marks of RichHF (Liang et al., 2024), which record where an artifact is, brush masks also record its extent. We convert each mask into a scalar artifact ratio by dividing the number of marked pixels by the total number of image pixels.

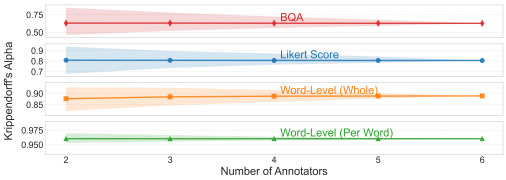
Figure 2b shows that brush-based annotation substantially improves reliability, increasing agreement to $\alpha = 0.82$ with CI $[0.56, 0.89]$. The confidence intervals are disjoint from those of Likert scoring, indicating that the gain is statistically meaningful at this sample size. The EDR also drops to 0%, showing that annotators make more consistent judgments about artifact presence and extent. Agreement heatmaps (Figure 12) show that annotators mark the same defect regions. The resulting masks also provide localized diagnostic feedback that may be useful for downstream model improvement (Liu et al., 2025; Fan et al., 2023; Wallace et al., 2024).

Convergence Analysis We further analyze how quickly each protocol stabilizes as annotators are added: for each annotator count k, we repeatedly sample random subsets of k annotators, compute Krippendorff's α, and report the mean and the 95% range across subsets as a shaded band. The band width measures how much the estimate depends on *which* annotators rate an item. As shown in Figure 2b, Likert scoring remains unstable even with six annotators and exhibits a wide band,



(a)

	Inter-Annotator Agreement (Krippendorff's α) ↑	EDR ($\geq 40\%$) ↓
BQA	0.63 [0.47, 0.75]	28.75%
Likert Score	0.81 [0.71, 0.86]	18.75%
Word-Level (Whole)	0.89 [0.79, 0.94]	**3.9%**
Word-Level (Per-Word)	**0.96** [0.95, 0.97]	4.5%



(b)

Figure 3: BQA and Likert (uniform baselines) vs. word-level annotation (ours) for text rendering: inter-annotator agreement (Krippendorff's α, 95% bootstrap CI), extreme disagreement rate, and convergence.

whereas brush-based annotation stabilizes with four annotators and keeps a narrow band across k, so the result does not depend on which annotators are sampled.

4.2 Text Rendering Accuracy

Earlier evaluations of text rendering often used BQA, asking whether the rendered text exactly matches the prompt (Cho et al., 2024); BQA and global Likert scoring are the uniform baselines here. This binary protocol becomes overly coarse when models produce mostly legible text with minor spelling or formatting errors.

To analyze this effect, we sample 30 prompts tagged with `text_rendering`, generate images using the same three models, and ask annotators under BQA whether the rendered text is correct. As shown in Figure 3b, BQA yields moderate agreement, with $\alpha = 0.63$ $[0.47, 0.75]$, and an EDR of 28.75%. Likert scoring improves agreement to $\alpha = 0.81$ $[0.71, 0.86]$ and reduces EDR to 18.75%. However, both BQA and Likert scoring compress text rendering into a single global judgment, discarding information about which words are correct, missing, or incorrectly inserted.

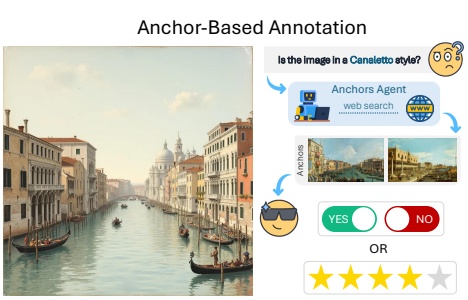
Skill-Aligned Protocol: Word-Level Annotation To align annotation with the discrete structure of text, we use word-level annotation, as shown in Figure 3a. Annotators judge each expected word individually as correct or incorrect. Missing words receive a negative judgment on the corresponding word tile. To capture insertion errors, we add virtual *gap tokens* between words and at sentence boundaries, which annotators mark when extra or repeated text appears. We compute the final accuracy as $\max(0, N_{\text{correct}} - N_{\text{incorrect_gaps}})/N_{\text{total_words}}$. This formulation penalizes incorrect words, missing words, and unintended insertions while preserving partial correctness. Partially rendered words count as incorrect; we chose word rather than character granularity to balance resolution against annotator effort.

With word-level annotation, agreement increases to $\alpha = 0.89$ $[0.79, 0.94]$, and EDR drops to 3.9%. When agreement is computed directly at the word and gap-token level, treating each tile as a binary item, agreement further increases to $\alpha = 0.96$ $[0.95, 0.97]$. These results show that decomposing text rendering into atomic units substantially reduces annotator variance compared to global judgments.

Convergence Analysis For BQA and Likert, agreement spans a wide band across annotator subsets, reflecting instability in global judgments, whereas word-level annotation shows a tight band.

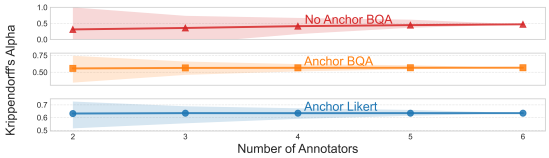
4.3 Skills Requiring External References

Some skills require external or domain-specific knowledge, such as recognizing landmarks, artistic styles, or named entities. If annotators are unfamiliar with the reference, their judgments may become unreliable or default to abstention. To investigate this, we sample 30 prompts from `environment:landmark`, `style:artistic_style`, and `named_entities`, and generate images using the same three models. As a baseline, annotators answer BQA questions with an explicit *Unsure* option, the uniform mechanism applied to these skills in prior work. As shown in



(a)

	Inter-Annotator Agreement (Krippendorff's α) ↑	Unsure Rate (%) ↓
No Anchor BQA	0.47 [0.25, 0.68]	63.0%
Anchor BQA	0.57 [0.47, 0.66]	**2.5%**
Anchor Likert	**0.64** [0.54, 0.72]	**2.5%**



(b)

Figure 4: Standard BQA (uniform baseline) vs. anchor-based protocols (ours) for reference-dependent skills: inter-annotator agreement (Krippendorff's α, 95% bootstrap CI), Unsure rate, and convergence.

Figure 4b, this baseline yields an *Unsure* rate of 63%, indicating that missing background knowledge is a major bottleneck.

Skill-Aligned Protocol: Anchor-Based Annotation To reduce dependence on annotator prior knowledge, we introduce *Anchor* annotation. Given a validation question, an automated retrieval step presents the top-k ($k = 3$) representative reference images for the queried entity or style alongside the generated image, and annotators consult these visual anchors before responding, as illustrated in Figure 4a.

Providing anchors reduces the *Unsure* rate to 2.5%, suggesting that most abstentions arise from missing reference knowledge rather than inherent ambiguity in the generated images. Agreement also increases from $\alpha = 0.47$ [0.25, 0.68] to $\alpha = 0.57$ [0.47, 0.66]. However, annotators reported that strict binary decisions can be too rigid for stylistic or referential similarity judgments. We therefore also evaluate *Anchor Likert*, which retains visual references but replaces binary responses with graded similarity ratings. Anchor Likert further improves agreement to $\alpha = 0.64$ [0.54, 0.72] while maintaining a low *Unsure* rate.

Convergence Analysis As shown in Figure 4b, standard BQA does not stabilize with six annotators, indicating persistent sensitivity to annotator choice. In contrast, Anchor BQA and Anchor Likert reach stable agreement with approximately four annotators and show narrower bands, *i.e.* less dependence on which annotators are sampled.

4.4 Discussion

Sample Size Rather than fixing the sample size in advance, we grew it until the reliability statistics converged. With 20 prompts per skill, the agreement estimates were still shifting and their confidence intervals overlapped across mechanisms. At 30 prompts (90 images per skill), the estimates stabilized, and the confidence intervals of the baseline and the skill-aligned protocol no longer overlapped for artifacts (Likert vs. brush) and text rendering (BQA vs. word-level), so additional images were not needed to separate the protocols.

Annotation Effort Both uniform and skill-aligned protocols begin with the same judgment phase, in which the annotator scans the whole image to locate artifacts or to check each rendered word; this phase is common to both and is where most of the effort lies. Uniform protocols then compress this judgment into a single Likert score or a yes/no answer, discarding the evidence the annotator has gathered. Skill-aligned protocols instead record it as artifact masks and per-word labels with only a few additional clicks, revealing *where* and *how* a model fails and directly supporting targeted mitigation, such as localized rewards for fine-tuning (Liu et al., 2025; Wallace et al., 2024). They also make better use of annotators, reaching stable agreement with four to five of them and reducing discarded *Unsure* answers from 63% to 2.5%, while all protocols run in a single interface with prompts routed automatically by the tagging step.

Model	Ent.	Attr.	Action	Spatial	Compar.	Light.	Weather	View	Camera
Nano-Banana	0.979	0.900	0.773	**0.978**	0.950	**0.955**	**1.000**	0.938	0.688
Wan-2.5-preview	**0.992**	0.920	0.793	0.960	**0.976**	0.875	**1.000**	**1.000**	0.778
FLUX.2-dev	0.934	**0.936**	**0.831**	0.880	0.743	0.907	0.933	0.688	0.882
Z-Image-Turbo	0.913	0.827	0.730	0.705	0.561	0.625	0.935	0.875	**0.905**

Model	Mood	Lang.	Time	Env.	Style	Named	Text	Artif.	Aesth.	Avg
Nano-Banana	**1.000**	**1.000**	0.967	0.901	**0.875**	0.639	0.930	**0.998**	**0.879**	**0.908**
Wan-2.5-preview	**1.000**	0.743	**0.969**	0.880	0.783	0.615	**0.980**	0.997	0.850	0.895
FLUX.2-dev	0.867	0.848	0.933	**0.908**	0.817	**0.659**	0.941	0.996	0.824	0.863
Z-Image-Turbo	**1.000**	0.176	0.767	0.835	0.691	0.485	0.877	0.995	0.796	0.761

Table 2: Per-skill comparison of diffusion models under the full skill-aligned evaluation protocol. Bold indicates best performance per skill. Abbreviations: **Ent.** Entities, **Attr.** Attributes, **Compar.** Relative Comparison, **Light.** Lighting, **Lang.** Language Complexity, **Env.** Environment/Landmark, **Artif.** Visual Artifacts, **Aesth.** Aesthetic Quality; per-skill sample sizes are in Appendix F.

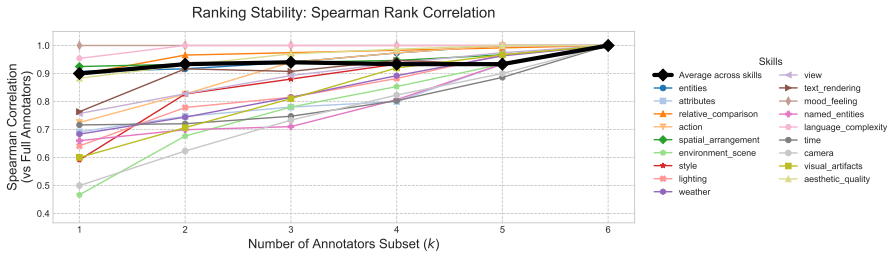


Figure 5: Score convergence under annotator subsampling. Spearman correlation between subset-based model scores and full-annotator scores is shown as a function of annotator count.

5 Full Evaluation Suite

Based on the controlled comparisons in the previous section, we assign each skill the annotation strategy with the strongest reliability profile: brush-based annotation for visual artifacts, word-level annotation with gap tokens for text rendering, Anchor Likert for reference-dependent skills, BQA for strictly factual skills, and Likert scoring for overall aesthetic quality.

For full evaluation, we sample 43 prompts from the taxonomy in Section 3.1, ensuring coverage across all skills and sub-skills (per-skill counts in Appendix F). We evaluate two proprietary models, Nano-Banana (Google DeepMind, 2025) and Wan-2.5-preview (WAN, 2025), and two open-source models, FLUX.2-dev (Black Forest Labs, 2025) and Z-Image-Turbo (Cai et al., 2025). To compare heterogeneous outputs, all non-binary scores, including Likert ratings, brush-based artifact ratios, and word-level accuracy, are normalized to $[0, 1]$. No existing uniform human protocol covers all 18 skills: BQA benchmarks cover semantic adherence only (Hu et al., 2023; Cho et al., 2024; Wiles et al., 2025), and preference studies do not decompose skills (Xu et al., 2023; Wu et al., 2023a), so a full-suite uniform baseline would require arbitrary mechanism-to-skill assignments. Section 4 therefore compares mechanisms per skill, where each comparison is controlled, and Section 6 compares the suite with VLM judges.

Table 2 summarizes per-skill performance across models. Nano-Banana obtains the highest average score, followed by Wan-2.5-preview, but aggregate rankings hide important skill-level differences. Brush-based artifact scores are nearly saturated across models, suggesting that recent T2I systems have largely closed the gap in low-level artifact suppression, whereas named entities remain challenging for every model, motivating stronger grounding mechanisms. Language complexity shows the largest variance: Z-Image-Turbo scores 0.176, far below the other systems, and is also weak on spatial arrangement and relative comparison, whereas Nano-Banana achieves perfect performance, consistent with stronger instruction following from its backbone LLM.

Model rankings are also stable under annotator subsampling: every skill exceeds a Spearman correlation of 0.8 with the full-annotator scores by four annotators (Figure 5; details in Appendix G).

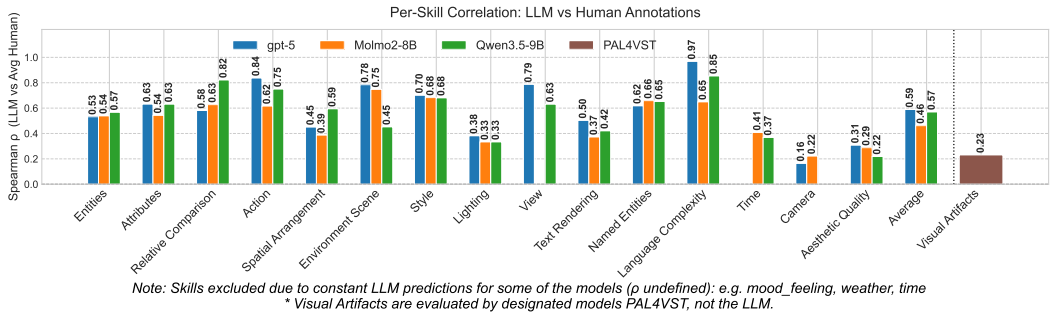


Figure 6: Spearman correlation between VLM-based automatic evaluation and human evaluation across skills. Structured semantic skills show higher alignment, while camera-related attributes and aesthetic quality exhibit lower correlation.

6 Automatic Evaluation with VLM Agents

To test fully automatic evaluation, we replicate our skill-aligned protocol with a lightweight agentic framework in which separate agents execute each skill-specific procedure. We evaluate two open-source VLMs, Molmo2-8B (Clark et al., 2026) and Qwen-3.5-9B (Qwen Team, 2026), and the proprietary ChatGPT-5. The BQA agent answers validation questions with *yes*/*no*; for reference-dependent skills, the agent receives a collage of the generated image and two labeled anchors and assigns a graded similarity score using the human Anchor Likert rubric. The word-level agent labels each expected word and gap individually, with accuracy computed as in the human protocol. Since VLMs cannot brush spatial artifacts, we use a pretrained artifact detector (Zhang et al., 2023) for artifact masks.

Figure 6 shows that most structured semantic skills achieve moderate to strong correlation with humans, and Qwen-3.5 performs comparably to ChatGPT-5 despite being orders of magnitude smaller. Aesthetic quality shows substantially lower alignment, suggesting generic VLM judgment is insufficient and may benefit from specialized predictors (*e.g.* ImageReward (Xu et al., 2023)), and camera-related attributes also correlate poorly. Artifact detection with PAL4VST shows relatively low agreement, likely because it was trained on earlier diffusion models. Among structured skills, `entities` and `spatial_arrangement` reveal limits in counting and multi-object reasoning, and word-level text rendering achieves only moderate alignment, indicating the need for a specialized model or human inspection.

7 Conclusion and Future Work

We studied *skill-aligned annotation* for text-to-image evaluation, where protocols are matched to the structure of each skill; in controlled comparisons, they produced more reliable and stable signals than uniform mechanisms while stabilizing with four to five annotators, and an automated pipeline instantiating them aligns strongly with human judgments on several skills.

Implications Per-skill scores localize failures rather than merely ranking models: Z-Image-Turbo's aggregate deficit is concentrated in language complexity (0.176) while its artifact suppression is on par with the best models, and named entities are the hardest skill for every model. Brush masks and per-word labels are dense signals that RL- or preference-based fine-tuning can consume directly (Liu et al., 2025; Fan et al., 2023; Wallace et al., 2024). Because the protocols stabilize with four to five annotators and are mirrored by the automated pipeline, models can be re-evaluated per skill at low cost, automating skills where VLM–human agreement is high and reserving human annotation for those where it is low. A skill's evaluation value is thus its discriminative power times its automation gap; end-to-end reward fine-tuning with these signals is future work.

Future Work Weather, camera, and mood/feeling also converge slowly across annotators (Figure 5, Figure 6) and warrant dedicated protocols. Brush annotation could also mark missing objects or parts, currently captured by BQA, and the methodology extends to axes such as safety, cultural alignment, and diversity. We hope this work encourages treating annotation design as a central component of evaluation methodology.

AI USE STATEMENT

Generative AI models are the object of study in this work: an LLM (GPT-5) performs skill tagging and validation-question generation (Section 3.1), and VLMs act as automatic evaluators in Section 6; both uses are described and evaluated in the paper. Beyond this, we used generative AI tools only for light editing of the text (grammar and phrasing) and for assistance with boilerplate code; no research ideas, experimental designs, results, or figures were generated by AI. All AI-assisted text and code were reviewed and verified by the authors, who take full responsibility for the content of this work.

ETHICS STATEMENT

This study involves human annotators who evaluated synthetic images generated by publicly available T2I models from publicly available prompts. Generated images may contain occasional visual artifacts, but no harmful, private, or offensive content was targeted. We release annotations, prompts, and tooling to support transparent and reproducible evaluation of T2I systems.

REPRODUCIBILITY STATEMENT

The annotation interface, task configurations, tagging prompt, automated evaluation agents, and all collected annotations are provided as supplementary material (Appendix B, Appendix C, and Appendix H). The skill taxonomy is specified in Table 1 and Table 4, the annotation protocols in Section 4, and the reliability metrics and sampling procedures in Sections 4.1 to 4.3. The prompt set is drawn from the public Gecko benchmark, and the evaluated models are publicly accessible.

REFERENCES

Black Forest Labs. FLUX.1. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-Image: An efficient image generation foundation model with single-stream diffusion Transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Jaemin Cho, Yushi Hu, Jason M. Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *ICLR*, 2024. URL https://openreview.net/forum?id=ITq4ZRUT4a.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026.

Abdelrahman Eldesokey, Aleksandar Cvejic, Bernard Ghanem, and Peter Wonka. Mind-the-Glitch: Visual correspondence for detecting inconsistencies in subject-driven generation. In *Advances in Neural Information Processing Systems*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow Transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=8OTPepXzeh.

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

Google DeepMind. Nano-Banana. https://deepmind.google/models/gemini-image/flash/, 2025.

Kevin David Hayes, Micah Goldblum, Vikash Sehwag, Gowthami Somepalli, Ashwinee Panda, and Tom Goldstein. FineGRAIN: Evaluating failure modes of text-to-image models with vision language model judges. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=qlZI9Bgxpy.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023.

Hang Hua, Ziyun Zeng, Yizhi Song, Yunlong Tang, Liu He, Daniel Aliaga, Wei Xiong, and Jiebo Luo. MMIG-Bench: Towards comprehensive and explainable evaluation of multi-modal image generation models. *arXiv preprint arXiv:2505.19415*, 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*, 2025.

Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, et al. LEGION: Learning to ground and explain for synthetic image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18937–18947, 2025.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, 2024.

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

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. GenAI-Bench: Evaluating and Improving Compositional Text-to-Visual Generation, November 2024.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19401–19411, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. LLMScore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in neural information processing systems*, 36:23075–23093, 2023.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14277–14286, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

Qwen Team. Qwen-3.5. https://qwen.ai/blog?id=qwen3.5, 2026.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Michael Saxon, Fatima Jahara, Mahsa Khoshnoodi, Yujie Lu, Aditya Sharma, and William Yang Wang. Who evaluates the evaluations? objectively scoring text-to-image prompt coherence metrics with t2IScorescore (TS2). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=S4YRCLbUK1.

Rong-Cheng Tu, Zi-Ao Ma, Tian Lan, Yuehao Zhao, He-Yan Huang, and Xian-Ling Mao. Automatic evaluation for text-to-image generation: Task-decomposed framework, distilled training, and meta-evaluation benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22340–22361, 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

WAN. WAN-2.5. https://wan.video/, 2025.

Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajic, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Pinelopi Papalampidi, Ira Ktena, Christopher Knutsen, Cyrus Rashtchian, Anant Nawalgaria, Jordi Pont-Tuset, and Aida Nematzadeh. Revisiting text-to-image evaluation with Gecko: on metrics, prompts, and human rating. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Im2neAMlre.

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

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2096–2105, 2023b.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. VisionReward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024.

Lingzhi Zhang, Zhengjie Xu, Connelly Barnes, Yuqian Zhou, Qing Liu, He Zhang, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. Perceptual artifacts localization for image synthesis tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7579–7590, 2023.

Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8018–8027, 2024.

Tianyu Zhang, Xinyu Wang, Lu Li, Zhenghan Tai, Jijun Chi, Jingrui Tian, Hailin He, and Suyuchen Wang. STRICT: Stress-test of rendering image containing text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21148–21161, 2025.

A Details on Skills Taxonomy

We provide additional examples illustrating the proposed skill–subskill taxonomy in Table 4, and an operational definition with inclusion and exclusion criteria for each skill in Table 3.

Construction The taxonomy was built in three steps. (1) *Seed.* We started from the Gecko taxonomy (Wiles et al., 2025), which covers prompt semantics such as entities, attributes, and spatial relations. (2) *Extension.* Since our goal is to evaluate all aspects of image generation, namely prompt adherence, perceptual quality, and aesthetics, we added the skills that semantics-focused taxonomies do not cover: visual artifacts and aesthetic quality. (3) *Iteration.* We tagged the Gecko prompts with the current taxonomy, manually reviewed the outputs, and extended or refined the taxonomy wherever an aspect of a prompt was not captured (an untagged phrase or a mis-assigned skill), then re-tagged. This loop was repeated until every prompt's evaluable content was covered, which is the stopping criterion. This loop added seven semantic skills absent from Gecko: lighting, weather, viewpoint, camera, mood/feeling, time (time of day, season, or era), and general environments and scenes. It also regrouped Gecko's skills: counting became part of *entities*, which gained uncountable entities; color, shape, texture, material, and scale became sub-skills of *attributes*; Gecko's comparative scale and quantifier prompts became *relative comparison*, which gained tone and distance; landmarks moved from named entities to *environment/scene*; and color Stroop prompts moved to *language complexity*.

Extensibility The framework remains extensible through three properties. First, adding a skill only requires editing the tagging prompt: the LLM tagger is instruction-based, so a new skill or sub-skill is introduced by adding its definition and an example, without retraining or pipeline changes. Second, protocol assignment follows the type of evidence a skill requires: a checkable fact maps to BQA, an image region to brush annotation, an external reference to anchor-based annotation, a token sequence to word-level annotation, and a holistic impression to Likert scoring, so new skills reuse the validated mechanisms. Third, the annotation interface is configuration-driven (Appendix C), so new skills slot in without modifying the application. Extending the suite to an axis such as safety or cultural alignment would thus amount to defining the skill in the tagging prompt, choosing its mechanism by evidence type (*e.g.* anchor-based for culture-specific references), and adding one entry to the task configuration.

Table 3: Operational definitions of the evaluation skills with inclusion and exclusion criteria.

Skill	Counts as (inclusion)	Does not count (exclusion)
Entities	An entity requested by the prompt: *Singular* (exactly one instance, "a cat"), *Count* (an explicit number, "three apples"), or *Uncountable* (a substance or mass noun, "milk"). A singular request fails if several instances appear.	Properties of the entity (Attributes); specific named instances (Named Entities); the background implied by an entity.
Attributes	A property stated of a specific entity: color, texture, material, shape, or scale ("a *wooden* chair").	Relative statements between two entities (Comparison); scene-level illumination or weather.
Action	What an agent is doing: *Standard* ("running"), *Unusual* (atypical for the agent, "a dog riding a skateboard"), or *Pose* (a static body configuration, "handstand").	Spatial relations between objects without an activity (Spatial Arrangement).
Spatial Arrangement	A stated spatial relation between two or more entities (on top of, left of, behind, inside).	Relative size or distance statements (Comparison).
Comparison	A relative statement between two entities: scale, tone, distance, count, or other ("more birds than airplanes").	Absolute properties of a single entity (Attributes).
Lighting	A stated illumination condition (soft studio light, backlit, neon glow).	Temporal cues that merely imply lighting ("sunset" is Time).
Weather	A stated atmospheric condition (snow, fog, rain).	Seasons (Time).
View	The viewpoint from which the scene is observed (top-down, side view, from below).	Lens and shot parameters (Camera).

Skill	Counts as (inclusion)	Does not count (exclusion)
Camera	Photographic parameters (macro, wide-angle, long exposure, shallow depth of field).	Viewpoint (View).
Mood / Feeling	An affective quality of the whole image (gloomy, cheerful, eerie).	Named artistic styles that imply a mood (Style).
Language Complexity	*Negation* (an element that must be absent, "no pedestrians") or *Color Stroop* (a color word to be rendered in a conflicting color).	Prompt length or syntactic complexity alone.
Time	A temporal setting: time of day, season, year, or era.	Illumination words without a temporal reference (Lighting).
Environment / Scene	*General* (a scene type such as a tropical beach or a kitchen) or *Landmark* (a specific named place).	Backgrounds only implied by the entities present.
Style	*Artistic Style* (a named artist, movement, or rendering style) or *Visual Medium* (the physical medium or surface, "painted on a vase").	Photographic parameters (Camera).
Named Entities	A specific real or fictional named instance: character, vehicle, product, or artwork.	Named places (Environment: Landmark); named artists (Style).
Text Rendering	Text to be rendered: *Accuracy* (the exact string), *Style* (font or script style), *Numerical* (digits and numeric expressions), or *Position* (placement relative to an object).	Color conflicts of the text (Language Complexity: Color Stroop).
Visual Artifacts	Localized rendering defects independent of the prompt: extra or fused limbs, warped geometry, texture noise, garbled glyphs. Assessed on every image.	Missing objects or parts, which are semantic failures captured by BQA.
Aesthetic Quality	Holistic visual appeal: composition, color harmony, sharpness. Assessed on every image.	Prompt adherence.

Table 4: Example prompts and validation questions for each evaluation skill and sub-skill. Questions follow the annotation strategies.

Skill	Sub-skill	Example Prompt	Validation Question
Entities	Singular	A photo of a cat	Is there exactly one cat in the image?
	Count	Three apples on a white plate	Are there three apples?
	Uncountable	A glass filled with milk	Does the image contain milk in the glass?
Attributes	Color	A blue bicycle	Is the bicycle blue?
	Texture	A rough stone sculpture	Does the sculpture have a rough texture?
	Material	A wooden chair in a living room	Is the chair made of wood?
	Shape	A perfectly round mirror hanging on a wall	Is the mirror round?
	Scale	A tiny elephant playing	Is the elephant tiny?
Action	Standard	A dog running across a field	Is the dog running?
	Unusual	A dog riding a skateboard	Is the dog riding a skateboard?
	Pose	A gymnast performing a handstand	Is the gymnast in a handstand pose?
Spatial Arrangement	–	A cup placed on top of a book	Is the cup on top of the book?
Comparison	Scale	A large elephant next to a small dog	Is the elephant larger than the dog?
	Tone	A darker cloud beside a lighter cloud	Is the left cloud darker than the right cloud?

Skill	Sub-skill	Example Prompt	Validation Question
	Distance	A tree closer to the camera than a house	Is the tree closer to the camera than the house?
	Count	More birds than airplanes in the sky	Are there more birds than airplanes?
Lighting	–	A portrait illuminated by soft studio lighting	Is the scene softly lit?
Weather	–	A snowy street during a snowstorm	Is it snowing in the scene?
View	–	A top-down view of a dining table with food	Is the scene viewed from above?
Camera	–	A macro close-up shot of a butterfly	Is this a close-up (macro) shot?
Mood / Feeling	–	A gloomy abandoned house under dark clouds	Does the image convey a gloomy mood?
Language Complexity	Negation	A street with cars but no pedestrians	Are there no pedestrians in the scene?
	Color Stroop	The word "RED" written in blue ink	Is the word RED written in blue color?
Time	Time of Day	A beach at sunset with orange sky	Is the scene at sunset?
	Season	Trees covered in autumn leaves	Does the scene depict autumn?
	Era	A medieval knight beside a castle	Does the scene depict a medieval era?
Environment / Scene	General	A tropical beach with palm trees	Does the image resemble a tropical beach scene?
	Landmark	The Eiffel Tower at night	Does the image depict the Eiffel Tower?
Style	Artistic Style	A portrait in the style of Van Gogh	How similar is the style to Van Gogh?
	Visual Medium	A landscape painting in on vase	Is the painting on a vase?
Named Entities	Character	Spider-Man swinging between buildings	Does the image depict Spider-Man?
	Vehicle	A Tesla Cybertruck parked on a street	Is the vehicle a Tesla Cybertruck?
	Product	A Coca-Cola can on a wooden table	Is the product Coca-Cola?
	Artwork	The Mona Lisa displayed in a museum	Does the image depict the Mona Lisa?
Text Rendering	Accuracy	A sign that reads "OPEN 24 HOURS"	Does the image have text "OPEN 24 HOURS"?
	Style	The word "Hello" in cursive style	Is the word "Hello" in cursive style??
	Numerical	A scoreboard showing 3–1	Does the scoreboard show numbers "3-1"?
	Position	A label "EXIT" above a door	Is the word "EXIT" above the door?

B SKILL TAGGING AGENT

We design a single agent that simultaneously performs three tasks: (1) tagging different parts of the prompt with the relevant skills, (2) generating validation questions, and (3) identifying dependency

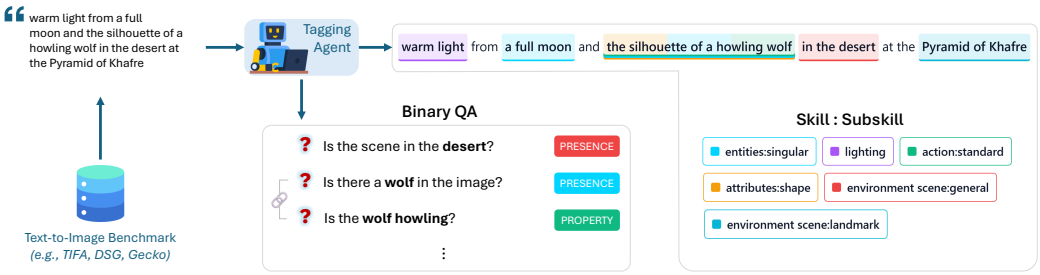


Figure 7: Our automated pipeline to tag Text-to-Image generation prompts with relevant evaluation skills and corresponding binary evaluation questions.

relationships between questions. While earlier approaches often relied on multiple agents or separate LLM calls for these steps, we find that recent LLMs can reliably perform them within a single prompt. In practice, generating the tags, questions, and dependencies jointly leads to more consistent and task-aligned outputs compared to performing each step independently. An overview of the tagging pipeline is shown in Figure 7. The full prompt is provided in the supplementary material. The tagger returns one record per skill–subskill pair with the following fields:

```
uid: str                # Unique ID per skill-subskill
skill: str              # Representative skill
subskill: str           # Representative subskill (empty if none)
phrase: str             # Phrase in the prompt representing the skill
question: str           # Binary QA validating the skill
node_type: str          # "presence", "property", or "relation"
depends_on: List[str]   # List of parent UIDs
```

Prompt Analysis Application To facilitate inspection of prompts and their automatically tagged skills and validation questions, we developed a lightweight web application. An overview of the interface is shown in Figure 8. The application summarizes the distribution of skills and sub-skills across all prompts in a dataset. When a specific prompt is selected, the interface displays detailed information including the tagged skills, the generated validation questions, and the dependency relationships between questions, as illustrated in Figure 9. A video demo of the application is provided in the supplementary material. The application and the tagged prompts are included in the supplementary material.

C Annotation Interface

We developed a web-based annotation interface to streamline the evaluation workflow. A screenshot of the interface is shown in Figure 10. The application loads tagged prompts according to a predefined task configuration specifying the annotation strategy for different skills. An example of task configurations is shown below:

```
{
  "id": "text_per_word",
  "name": "Text Per Word",
  "enable_bqa_ai": false,
  "shuffle_images": true,
  "annotations": [
    "text_per_word"
  ],
  "dataset_version": "v8.1-gpt-5-mini",
  "prompts_file": "text_rendering_collection.json",
  "models": [
    "z-image",
    "flux2-dev",
    "flux1-dev"
  ]
}
```

Generated images are then displayed together with their evaluation questions and corresponding annotation controls. For Anchor-based annotations, an image icon appears next to the question and reveals the anchor reference images when hovered over. If a child question depends on a parent question, the child is disabled if the answer to the parent is negative.

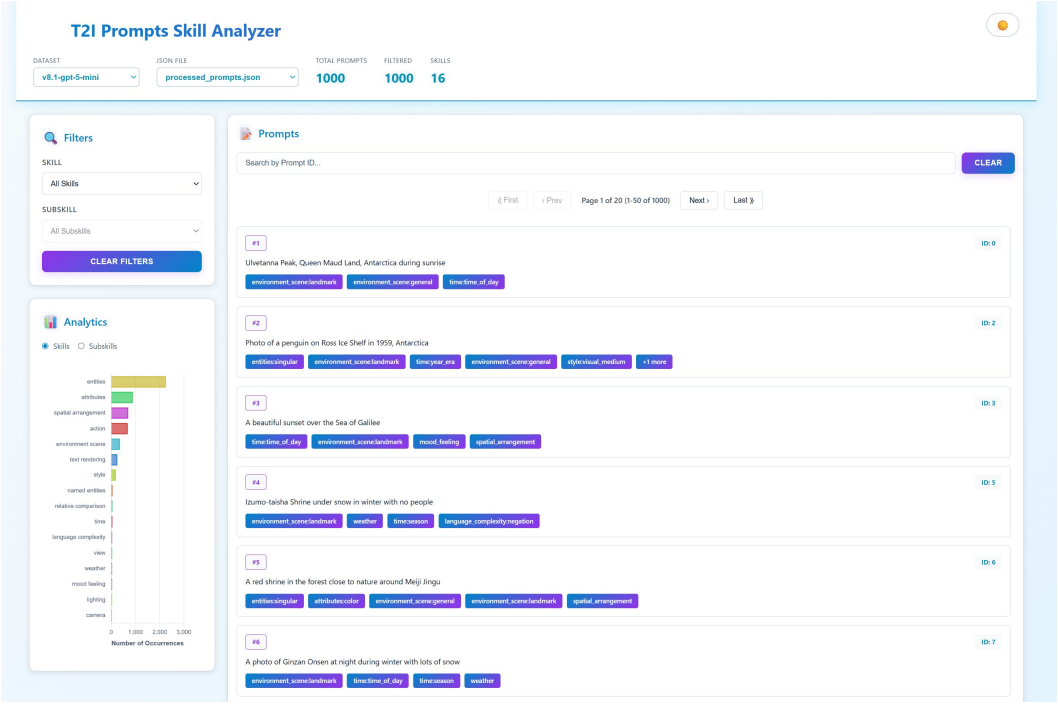


Figure 8: Overview of the prompt analysis interface. The application summarizes the distribution of skills and sub-skills across the prompts of a dataset and allows users to browse individual prompts.

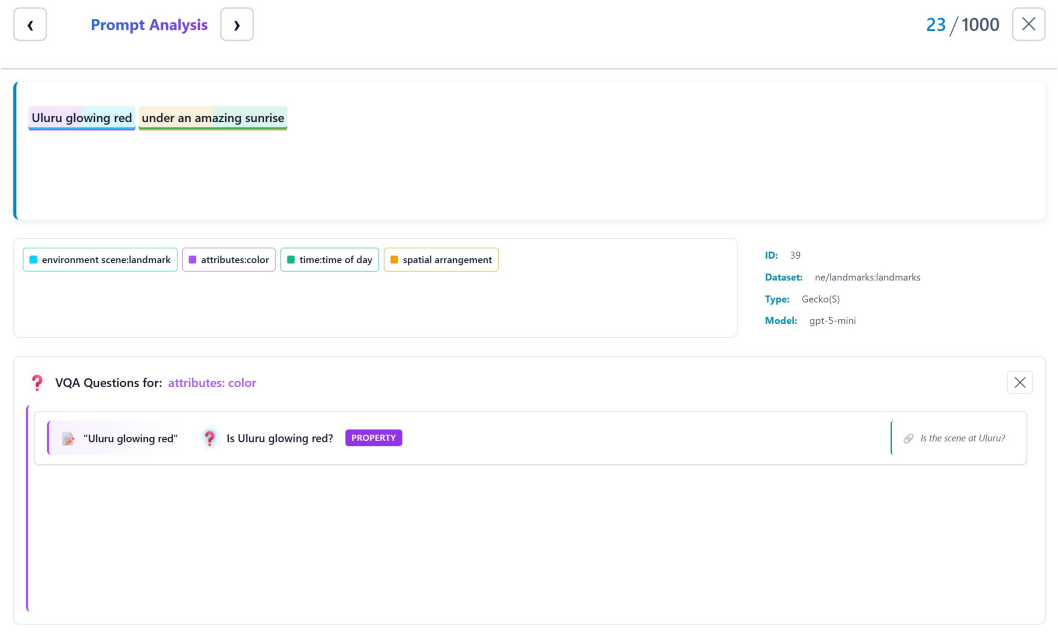


Figure 9: Detailed view for a selected prompt showing the automatically tagged skills, generated validation questions, and dependency relationships between questions.

To accelerate annotation of the full suite (Section 5), BQA answers were initialized by a VLM and shown as suggestions explicitly labeled *AI*, which annotators confirmed or corrected; no suggestions were used for word-level, anchor, brush, or aesthetics annotations, nor in any of the protocol comparisons of Section 4. For artifact annotation, users can mark defective regions directly on the image. When selecting the artifact tool, the image becomes a drawing canvas that allows annotators to brush over areas containing visual artifacts, as illustrated in Figure 11. A video demo of the

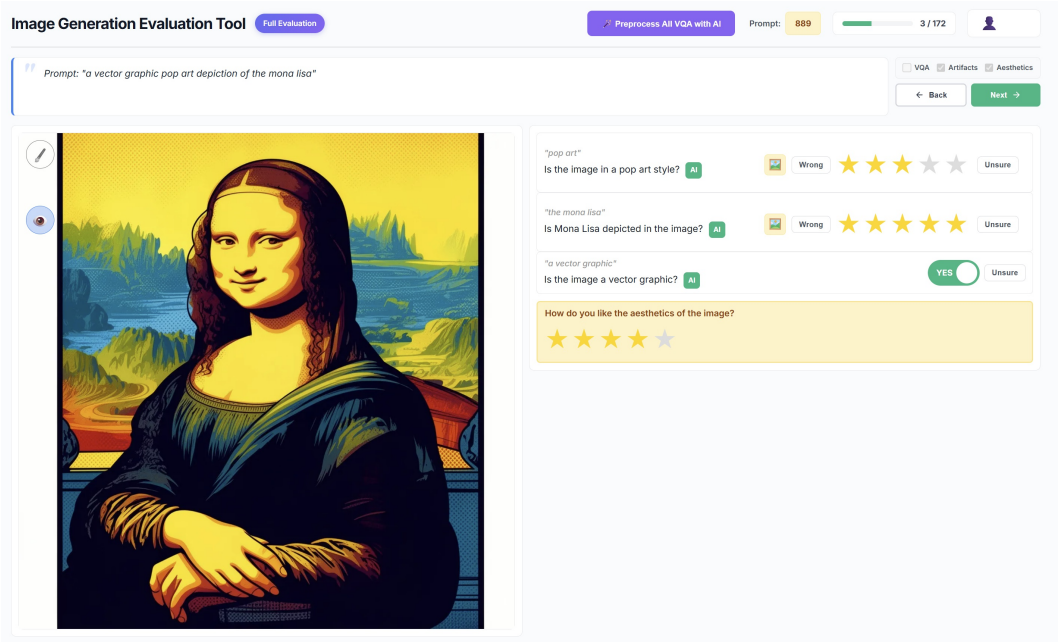


Figure 10: Web-based annotation interface used for skill-specific evaluation. Prompts, generated images, and corresponding annotation controls are displayed together according to the configured evaluation protocol.

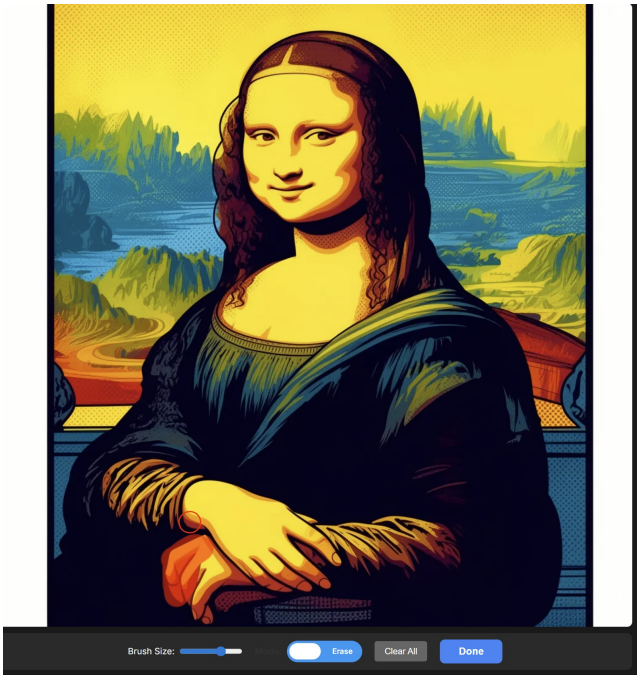


Figure 11: Artifact annotation tool. Annotators can brush over image regions that contain visual artifacts, enabling spatially localized artifact scoring.

annotation app is provided in the supplementary material. The source code of the application is included in the supplementary material.

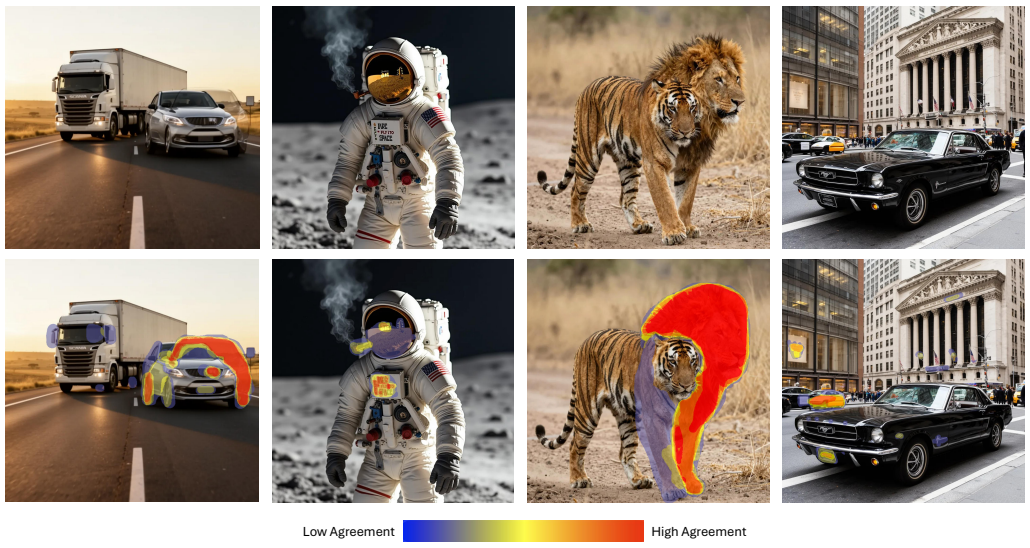


Figure 12: Examples of brush-based artifact annotations with inter-annotator agreement visualized as heatmaps. Red indicates strong agreement, while yellow and blue correspond to progressively weaker agreement.

D Annotation Examples

Artifacts Brush Figure 12 illustrates examples of brush-based artifact annotations together with the resulting inter-annotator agreement heatmaps. Regions with strong agreement (red) correspond to clear visual defects where annotators consistently identify artifacts. Areas with moderate agreement (yellow) typically reflect differences in how extensively annotators mark the spread of an artifact, which depends on how thoroughly the brush is applied. Finally, regions with low agreement (blue) indicate more subjective cases where some annotators label phenomena such as compression artifacts or reflections while others do not. These ambiguities could be reduced with more detailed annotation guidelines. However, we intentionally kept the instructions minimal to mimic realistic crowd-sourced evaluation settings commonly used in text-to-image benchmarking.

Word-Level Text Figure 13 presents examples of word-level annotations for evaluating text rendering accuracy. Unlike standard word-level annotation, which cannot capture extra words inserted between requested tokens, our scheme explicitly includes *gaps* between words. This allows annotators to identify spurious text that commonly appears in T2I outputs. The word-level formulation also provides localized feedback by identifying which specific words fail to render correctly. This is particularly useful in challenging cases where correct text overlaps with gibberish, such as the dragon example in Figure 13. In such cases, automated approaches often fail to decode the text reliably, whereas human annotators can still accurately verify individual words.

Anchor Likert Figure 14 presents examples of anchor images used for different skills. Directly answering some evaluation questions can be difficult for annotators who are not familiar with the subject. Providing reference anchors converts the task into a visual comparison problem, making it easier to judge the similarity between the generated image and known reference examples.

E EDR Threshold Sensitivity

The extreme disagreement rate (EDR) of Section 4.1 uses a threshold τ of 40% of the scoring range, *i.e.* two points on a 1–5 Likert scale. The value was not tuned. A one-point difference (*e.g.* 4 vs. 5) is ordinary calibration noise between annotators, and 98.9% of artifact images show at least a one-point difference, so $\tau = 1$ would flag nearly every image and carry no information. A two-point gap (*e.g.* "good" vs. "poor") marks genuinely different verdicts. The same 40%-of-range rule is applied to continuous scores such as artifact ratios and word accuracy. Table 5 recomputes EDR at 20%, 40%, and 60% of the range: at every threshold the skill-aligned protocol shows equal or lower extreme

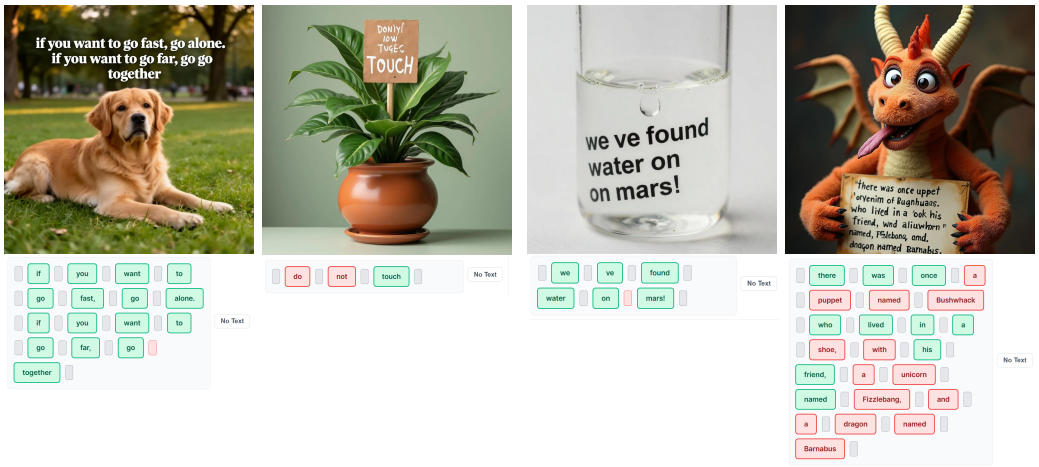


Figure 13: Examples of word-level text annotations illustrating correct words, missing words, and spurious insertions captured through gap annotations.

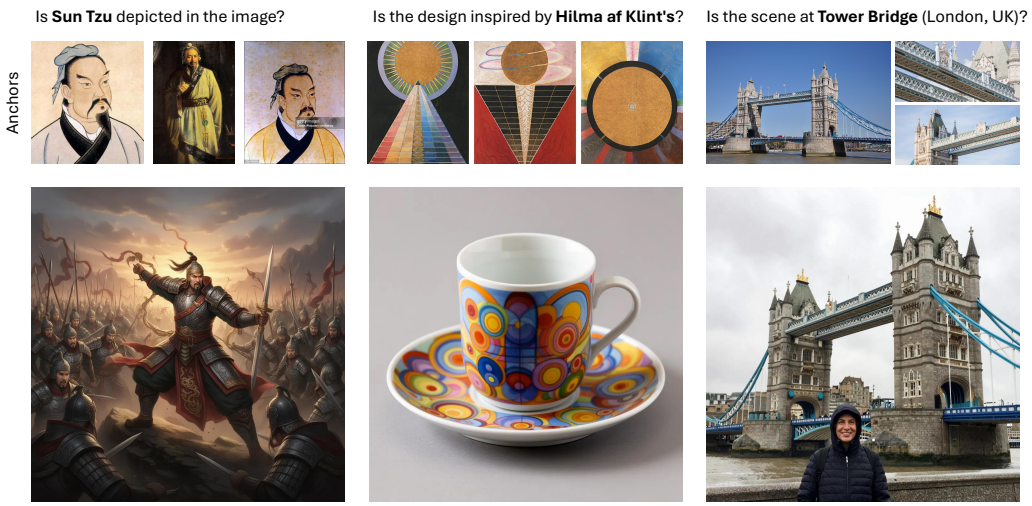


Figure 14: Examples of anchor references used to guide BQA or Likert scoring across categories such as characters, artwork styles, and landmarks.

disagreement than the uniform baseline, and brush annotation stays at 0% even under a much stricter threshold of 20 percentage points of artifact area (5.6% at 5 points).

Skill	**Protocol**	$\tau = 20\%$	$\tau = 40\%$ (paper)	$\tau = 60\%$
Visual artifacts	Global Likert (baseline)	98.9%	62.2%	20.0%
	Brush (ours)	0.0%	0.0%	0.0%
Text rendering	Likert (baseline)	57.8%	18.75%	1.2%
	Word-level (ours)	18.3%	3.9%	1.2%

Table 5: Extreme disagreement rate as a function of the threshold τ, expressed as a fraction of the scoring range. The ordering of protocols is unchanged at every threshold.

F Skill Distribution of the Full Evaluation Suite

Table 6 lists, for the 43-prompt suite of Section 5, the number of prompts that trigger each skill and the number of validation questions. All 16 semantic skills are covered with 294 questions in total; visual artifacts and aesthetic quality are assessed on every image and are therefore not part of the

question-based distribution. The distribution mirrors the natural frequency of skills in the Gecko corpus from which the prompts are sampled: compositional skills such as entities appear in most prompts, whereas camera and view are inherently infrequent. Because the protocol comparisons of Section 4 are run on dedicated single-skill prompt sets, this imbalance does not affect the reported reliability gains.

Skill	#prompts	#questions	Skill	#prompts	#questions
Entities	35	84	Named entities	7	7
Attributes	18	43	Time	7	8
Spatial arrangement	15	23	Relative comparison	6	7
Action	14	26	Language complexity	6	6
Environment / scene	14	22	Lighting	6	10
Style	12	16	Mood / feeling	4	4
Text rendering	11	23	View	3	3
Weather	7	8	Camera	3	4

Table 6: Number of prompts (of 43) and validation questions per semantic skill in the full evaluation suite; 294 questions in total.

G Score Convergence Across Skills

We evaluate score stability by measuring how model rankings change under annotator subsampling. For each annotator count k, we sample 50 random subsets of k annotators, recompute model scores, and measure the Spearman rank correlation with the full-annotator scores. Figure 5 in Section 5 shows that rankings remain stable across annotator counts, and aggregate correlations are consistently high, indicating that the overall model ordering is robust to annotator subsampling. At the skill level, convergence varies slightly: `relative_comparison` and `language_complexity` stabilize rapidly, whereas `environment_scene` and `style` require more annotators due to greater subjectivity and reliance on reference anchors. Nevertheless, all skills exceed a Spearman correlation of 0.8 by four annotators, showing that the evaluation suite yields stable model rankings with a modest annotator pool.

H Automated Evaluation Pipeline Details

The automated evaluation pipeline mirrors the human protocol using a set of specialized agents, each implementing the annotation strategy associated with a particular skill. In total, the system includes four agents: Binary QA, Anchor Likert, Aesthetics Scoring, and Artifact Detection. The open-source VLMs (Molmo2-8B and Qwen-3.5-9B) were deployed on a single A100 GPU (80GB VRAM).

(1) Binary QA Agent This agent evaluates factual skills using binary question answering (BQA), such as object presence, counting, and spatial relations. For each image, a batch of validation questions is provided together with unique identifiers (UIDs). The agent must respond to each question with `yes`, `no`, or `unsure`. To avoid ambiguity in object counting, the prompt explicitly instructs the model that if a prompt specifies *a/an* object but the image contains multiple instances, the answer should be considered incorrect. The system prompt is shown below.

```
# Answer the following questions about the image with only 'yes', 'no', or 'unsure'.
# For questions about presence of objects, if the prompt says a/an <object>,
# but the image has more, this should be flagged as wrong.
# UID: {question_uid_1}  Question: {question_text_1}
# UID: {question_uid_2}  Question: {question_text_2}
# ...
```

The output is programmatically validated to ensure that each UID is present and that the responses contain only valid labels.

(2) Anchor Likert Agent This agent evaluates reference-dependent skills using the Anchor Likert protocol. The model receives a visual grid containing the generated image and two reference anchors that represent canonical examples of the evaluated concept. The agent is instructed to rate the similarity of the generated image to the reference examples on a scale from 0 to 5, following the same rubric used for human annotators.

0.5884	0.5830	0.5397	0.6662	0.6989	0.5743	0.8040	0.3158	1.0000
								Avg.
0.6455	0.8442	1.0000	0.7018	1.0000	0.4112	0.1026	0.2399	0.6303

Table 7: Inter-VLM agreement (Krippendorff's α) per skill. VLMs exhibit moderate to high agreement across most skills except for lighting, camera, and aesthetic quality. Icons correspond to: Entities, Attributes, Action, Spatial Arrangement, Relative Comparison, Lighting, Weather, Viewpoint, Camera Control, Mood/Feeling, Language Complexity, Time, Environment/Landmark, Style, Named Entities, Text Rendering, and Aesthetic Quality.

```
# You are evaluating a generated image against reference images.
# The leftmost image is the generated image. The other image(s) are reference examples.
# Question UID: {uid}
# Question: {question}
# Phrase being evaluated: "{phrase}"

# Rate how well the generated image matches the reference
# on a scale from 0 to 5:
# 0 = completely wrong / not present
# 1 = barely recognizable
# 2 = somewhat resembles
# 3 = moderately accurate
# 4 = mostly accurate
# 5 = perfectly matches
# Or answer 'unsure' if you cannot determine.
```

(3) Aesthetics Score Agent This agent assesses the overall visual appeal of the generated image. The model is asked to rate the image on a 1–5 Likert scale reflecting perceived aesthetic quality, visual coherence, and attractiveness.

```
# Rate the overall aesthetics, visual quality, and attractiveness of this image on a scale
    from 1 to 5:
# 1 = Very poor quality, unattractive, or heavily distorted
# 2 = Poor quality, noticeable flaws or unattractive
# 3 = Average quality, acceptable but not particularly impressive
# 4 = Good quality, visually pleasing with minor to no flaws
# 5 = Excellent quality, highly attractive and visually striking
```

(4) Artifact Detection Agent Spatial artifact annotation cannot be directly produced by VLMs. Instead, we rely on pretrained artifact detection models to approximate the brush-based human annotations. In particular, we employ PAL4VST (Zhang et al., 2023), which produces pixel-level artifact masks that are used to compute artifact scores via mask area ratios. We also experimented with LEGION (Kang et al., 2025). However, we observed that it often highlights entire objects containing artifacts rather than localizing the artifact regions themselves. This behavior is incompatible with our definition of artifacts, which focuses on localized visual defects, and therefore PAL4VST was used in the final pipeline.

I Inter-VLM Agreement

To further examine the consistency of the protocol across models, we measure inter-VLM agreement using Krippendorff's α. The results are summarized in Table 7. Agreement is high for most structured skills, suggesting that different VLMs tend to produce consistent annotations under the proposed evaluation framework. In contrast, camera-related attributes, lighting, and aesthetic quality show lower agreement, reinforcing the observation that these skills remain challenging for automatic evaluation.



Figure 15: Examples of challenging annotation scenarios. (a) Overlapping chair legs create ambiguity in artifact localization. (b) Style matching remains difficult for non-expert annotators even with anchor references.

J CHALLENGING SCENARIOS

During our analysis, we observed several challenging cases for human annotation, illustrated in Figure 15. In the chair example (a), the legs overlap heavily, making it difficult for annotators to determine which regions should be marked as artifacts. Despite this ambiguity, the agreement heatmap still reveals regions of strong consensus where the visual defect is clearly identifiable.

In the style example (b), even with anchor references, annotators without an art background may struggle to determine whether the generated image matches the target style. This difficulty arises because recognizing stylistic similarity often requires attention to specific attributes such as brush strokes, color intensity, or composition. These challenges could be mitigated with clearer annotation guidelines or lightweight domain hints that help annotators focus on the relevant visual cues. Exploring such strategies is an interesting direction for improving annotation reliability in future work.