Imag-Eval: a language-grounded framework for interpretable Text-to-Image instruction following evaluation

Ibrahim MOHAMED SEROUIS *
Talan Research and Innovation Center
Toulouse, France
ibrahim.mohamed-serouis@talan.com

David JARAMILLO DUQUE
Talan Research and Innovation Center
Paris, France
david.jaramillo-duque@talan.com

Abstract

Text-to-Image (T2I) models have recently achieved impressive visual fidelity, yet their evaluation remains constrained by benchmarks that are often difficult to interpret and insufficiently diagnostic. Existing skill-based evaluations tend to overlook critical failure modes that strongly impact usability but fall outside standard taxonomies, such as global incoherence arising from missing parts or physically implausible configurations (e.g., floating objects). In addition, prompt difficulty is typically controlled along a single dimension; either prompt length or the number of elements to generate. To address these limitations, we introduce IMAG-EVAL, a controlled benchmark designed to assess how T2I models ground compositional natural-language instructions into visual outputs. Unlike prior work that conflates surface linguistic complexity with compositional difficulty, IMAG-EVAL explicitly seeks to disentangle these factors by independently varying both the number of instances and the combination of constraints (rules), while avoiding error propagation. This design enables fine-grained and interpretable analysis of *where* cross-modal instruction following fails. Our benchmark comprises 1,140 prompts and 8,842 combined rules, and we evaluate it on several state-of-the-art models. Complementing this analysis with an additional study of over 2,000 prompts from a concurrent benchmark, our results suggest that, for structured skills, compositional difficulty is primarily governed by the number of grounded rules and their binding to instances, rather than by prompt length alone. Code samples are available here: https://github.com/Justsecret123/Imag-Eval

1 Introduction

Text-to-Image generation has advanced rapidly, with modern systems capable of producing visually realistic and diverse images from natural language prompts, spanning a wide range of styles (Yu et al., 2024), lighting conditions, and even multilingual rendering. Despite these improvements, reliable and interpretable evaluation remains a bottleneck.

Early evaluation approaches relied on embedding-based metrics such as CLIPScore (Hessel et al., 2021), which approximate text–image alignment but provide limited diagnostic insight. More recent benchmarks move toward skill-based evaluation, decomposing performance into capabilities across different difficulty levels. However, these approaches remain limited in interpretability. First, they often fail to capture whether all required entities are generated coherently, overlooking critical failure modes such as missing parts or globally inconsistent scenes. Second, prompt difficulty is typically controlled along a single axis, most commonly prompt length, or per-skill hardness without explicitly accounting for the underlying compositional structure of the instruction. Third, evaluation pipelines frequently propagate upstream errors (e.g., missing objects) into downstream skill failures, leading to double penalization and obscuring the source of model breakdowns. To address these limitations, we introduce:

(1) Imag-Eval, a controlled evaluation framework designed to provide fine-grained and interpretable analysis of compositional instruction following in T2I models. Unlike prior approaches that rely on proxy measures such as prompt length, Imag-Eval explicitly seeks to disentangle surface linguistic complexity from compositional difficulty by varying two orthogonal factors that we name *compositional load*: the number of instances and the combination of constraints. This design enables us to directly investigate where models fail as compositional load increases. Furthermore, we introduce an under-explored yet critical evaluation dimension, coherent and complete object generation.

(2) A benchmark dataset supporting this frame-

*Corresponding author, main contributor.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 29236–29254
October 24-29, 2026 ©2026 Association for Computational Linguistics

work, comprising 1,140 prompts and 8,842 evaluation rules. We validate the proposed methodology through extensive experiments involving 14 annotators, over 8,000 annotations, and more than 6,000 generated images evaluated across a diverse set of proprietary and open-source models.

Our empirical findings, including an additional analysis of more than 2,000 prompts from a concurrent benchmark, provide broader insights into current evaluation practices. In particular, we show that model performance is primarily driven by *compositional load* (the number of grounded rules and their binding to instances) rather than by surface linguistic properties such as prompt length, for structured skills. This highlights the importance of disentangling linguistic complexity from compositional structure, and motivates the need for controlled benchmarks that enable more interpretable and diagnostic assessment of T2I models.

2 Related Work

Interpretable/Explainable AI (XAI) aims to make model behavior and failure modes understandable, enabling more reliable diagnosis and comparison beyond aggregate metrics (Ribeiro et al., 2016; Doshi-Velez and Kim, 2017; Lipton, 2018). A key principle is that evaluation should be *diagnostic*: rather than reporting a single scalar score, it should decompose performance into meaningful dimensions that help localize errors. This requirement is particularly critical for generative multimodal systems, where high-level similarity metrics (e.g., CLIP-based scores (Hessel et al., 2021)) can conflate distinct failure sources, such as missing entities and incorrect bindings. In response, recent work advocates skill-based evaluations that assess models along semantically grounded axes (Ribeiro et al., 2020; Srivastava et al., 2023; Afkanpour et al., 2025).

Image generation. Image generation has evolved from generative adversarial networks (GANs) (Goodfellow et al., 2014) and variational autoencoders (VAEs) (Kingma and Welling, 2013) to diffusion-based models (Sohl-Dickstein et al., 2015; Ho et al., 2020), which now dominate the field. Diffusion models achieve strong fidelity and diversity through iterative denoising and underpin many modern T2I systems (Rombach et al., 2022). More recent architectures (e.g., Stable Diffusion XL, Stable Cascade, Qwen-Image) further extend these capabilities, improving both visual quality and controllability (Podell et al., 2024; Pernias et al., 2024; Wu et al., 2025). Despite these advances, evaluation has not kept pace, particularly in compositional instruction following.

Text-to-Image benchmarks. Early multi-task T2I benchmarks defined broad evaluation categories, sometimes with difficulty tiers, but typically assessed skills in isolation, limiting the study of *compositional* failures across interacting constraints (Petsiuk et al., 2022). DALL-Eval (Cho et al., 2023) introduced automated evaluation for a limited set of skills (e.g., counting, spatial relations), but provides limited coverage of global coherence issues. More recent benchmarks incorporate multiple constraints within single prompts. HRS-Bench (Bakr et al., 2023) expands coverage to diverse attributes, including emotion and robustness, yet its notion of difficulty is primarily rule-centric. Multi-skill evaluations such as T2I-CompBench(++) (Huang et al., 2023, 2025) assess attribute binding and object relations across curated tasks, but do not explicitly control difficulty across multiple interacting factors.

TIIF-Bench (Wei et al., 2025) highlights the role of prompt length by comparing equivalent short and long prompts, demonstrating that length can confound evaluation. However, as suggested in our analysis, prompt length alone is an incomplete proxy for difficulty, as it may vary independently of the underlying compositional structure. This observation motivates the need to disentangle surface linguistic properties from compositional load when assessing instruction-following capabilities.

Across benchmarks, an additional limitation is that evaluation pipelines often propagate upstream errors (e.g., missing objects) into downstream skill failures, effectively double-counting errors and obscuring their origin. Moreover, global coherence issues such as missing parts or physically implausible configurations remain underrepresented in standard taxonomies. While specialized efforts target related artifacts (e.g., human-body realism), they do not provide a general, model-agnostic framework for evaluating coherence across all entities (Andreou et al., 2024; Corneanu et al., 2025).

Positioning of Imag-Eval. IMAG-EVAL builds on these lines of work by introducing a controlled and explicitly factorized evaluation framework. In contrast to prior benchmarks that rely on single-axis proxies, our approach disentangles surface linguistic complexity from compositional load by independently varying the number of instances and

the combination of constraints. This design enables fine-grained, interpretable analysis of cross-modal grounding under increasing compositional complexity, while isolating failure modes without confounding error propagation. As such, it complements existing benchmarks by providing a diagnostic perspective centered on compositional structure.

3 Proposed method: Imag-Eval

3.1 Skills definition

3.1.1 Common evaluation skills

(1) Counting measures how accurately the model generates the requested number of object instances. For example, if the prompt specifies five apples, we check whether exactly five are present. Each object is scored as 0 (failure) or 1 (success), and the final score is the ratio of successful cases to the total number of objects. This is the only skill evaluated at the object level rather than the instance level. An example : `{"object": "apple", "count": 5}`.

(2) Spatial Relationships assesses whether the model correctly positions objects relative to one another (e.g., "apple under table"). The score reflects the average success rate across all instances. An example rule: `["apple", "under", "table"]`.

(3) Color checks adherence to assigned colors for each object instance. When a color is specified for an object, all its instances must comply. It is embedded within the Counting rule, e.g., `{"object": "apple", "count": 5, "color": "blue"}`.

(4) Size verifies compliance with relative size relationships between instances. An example rule: `["apple", "smaller", "table"]`.

(5) Emotion evaluates how well the model conveys specified emotions for characters. Success is determined per instance, based on whether the generated emotion matches the prompt. An example rule: `"a person next to the fire hydrant is rejoicing"`.

(6) Text measures the accuracy of generated text against the prompt. Success is evaluated based on how closely the generated text matches the required text. An example : `A small sign on the wall reads: "Toilet, hot dog: not for sharing"`.

3.1.2 Under-explored evaluation skill

(7) Cohesiveness. Determines whether all generated instances are complete and coherent. This binary skill is scored as False if any object lacks essential parts (e.g., a headless human) unless explicitly requested, and true otherwise. As illustrated in Fig.1, an image can satisfy multiple criteria, but remain unusable in most contexts due to an absence of global *Cohesiveness*.

Figure 1: *Cohesiveness*. Dall-E positioned the elements as required, respected spatial relationships, generated the text with the required typo, but created a human instance with an extra arm.

We define a lack of *Cohesiveness* as the occurrence of anatomical or structural inconsistencies. These include unprompted deformed, missing, or extra body parts, incomplete objects or surfaces, and physically implausible arrangements (e.g., floating objects).

3.2 Compositional load

Our structure involving multiple skill combinations enables controlled evaluation at multiple perspectives: from simpler settings requiring only a small number of skills (two-skill combinations) to more demanding configurations that combine several skills (up to 6-skill combination), with or without robustness testing. We assign up to 5 instances in **hard** examples, up to 3 instances for **medium**, and up to 2 in **easy** examples. We define *compositional load* as the number of grounded constraints and their binding to instances. Operationally, we compute the compositional load as the sum of the number of instantiated objects and the number of constraints that must be simultaneously satisfied. Formally, for a prompt p, the compositional load $CL(p)$ is defined as:

$$\begin{aligned} CL(p) = I(p) + E(p) + T(p) + \\ C_w(p) + Si_w(p) + Sp_w(p) \end{aligned} \tag{1}$$

where $I(p)$ is the number of object instances, $E(p)$ the number of emotion constraints, $T(p)$ the number of text constraints, and $C_w(p)$, $Si_w(p)$, and $Sp_w(p)$ the instance-weighted counts of color, size, and spatial constraints.

3.3 Dataset

3.3.1 Prompt generation steps

(1) Meta-prompt initialization. We begin by constructing a meta-prompt, prefixed with the instruction: *Create a natural language text description for an image that contains the following elements*. This meta-prompt serves as the foundation for generating the final prompt, named *synthetic prompt*.

(2) Rule composition. For each skill under evaluation, we append the relevant parameters to the meta-prompt. For example, when testing *Counting+Color*, we include specifications such as `{"object": name, "count": object_count, "color": assigned_color}`, with object names randomly drawn from the COCO dataset (Lin et al., 2014). For the *Emotion* skill, emotions are assigned exclusively to human figures within the scene, ensuring each emotion is tied to a single individual. For *Size* and *Spatial Relationships*, we append relational instructions (e.g., `{A, smaller, B}` or `{A, under, B}`) after the object declarations to maintain logical consistency. For the *Text* skill, we include a textual element (e.g., a sign) that references specific objects from the prompt.

(3) Object assignment. We then assign object counts based on difficulty level: up to 5 objects per skill (3 for emotion) in hard examples, up to 3 objects (2 for emotion) in medium examples, and up to 2 objects (1 for emotion) in easy examples. This assignment is streamlined by our JSON-based skill architecture, which we convert into a string and append to the meta-prompt.

(4) Semi-modular prompt generation. Finally, we use a text generation model to produce a description of the hypothetical image. To increase robustness in challenging cases, we create a version of each test with lexical and semantic perturbations, by introducing typos and synonyms. For each sample, the meta-prompt is dynamically constructed based on the skills being evaluated. With this approach, we generated a dataset of 1,140 prompts, manually verified with respect to the JSON rules. Dataset statistics are available in App.C.

Fig 2 illustrates an example of a synthetic prompt generated from such a meta-prompt, specifically for a test combining Color, Emotion, and Text skills at an easy level. This example highlights how elements from the meta-prompt are naturally integrated into the final text description. To generate these text descriptions, we selected models from the literature that demonstrated strong comprehension abilities, specifically those achieving over 80% accuracy on the MMLU-Pro benchmark (Hendrycks et al., 2021). This benchmark's focus on understanding capabilities made it particularly relevant for our needs, as it directly relates to a model's ability to understand the context. We also considered resource constraints and model availability during our selection process. Focusing on recent models (2024 and later), we generated 30 sample synthetic prompts using local Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025), Mindlink-32B, Deepseek R1 (Guo et al., 2025), and GPT-5 (Singh et al., 2025). After comparing the outputs, we selected GPT-5, as its resulting prompts demonstrated fewer errors and greater originality. **However, our JSON rule-based structure allows for generating with any other model. Minor tweaks can also be applied to the code to increase the number of levels, instances, or rules, by modifying a parameter within the scripts. A description on the "how" is available in App.A.**

It is important to note that proprietary LLM APIs and their parameters may evolve over time, which can introduce minor variations in prompt generation. Nevertheless, our experiments indicate that the relative differences in compositional load are generally preserved across such variations.

```
A street scene with a bright yellow
parking meter in the foreground and
a large pink bear standing on the
sidewalk; a surprised person nearby
with mouth agape and hands raised is
clearly reacting to the bear; a sign
on a lamppost prominently reads "Pink
Bear Ahead".
```

Figure 2: Prompt example : *Color+Text+Emotion* skills. Highlighted parts are derived from our JSON rules.

3.3.2 Image generation

For image generation, we evaluated several models from the literature, considering factors such as parameter count, VRAM consumption, the balance between open-source and proprietary models, and publication year. Given the rapid advancements in image generation over the past three years, we excluded models whose latest versions were released before 2022. We prioritized models supported by whitepapers or peer-reviewed studies, as well as those developed by reputable sources. Fig.3 shows a generated sample from the synthetic

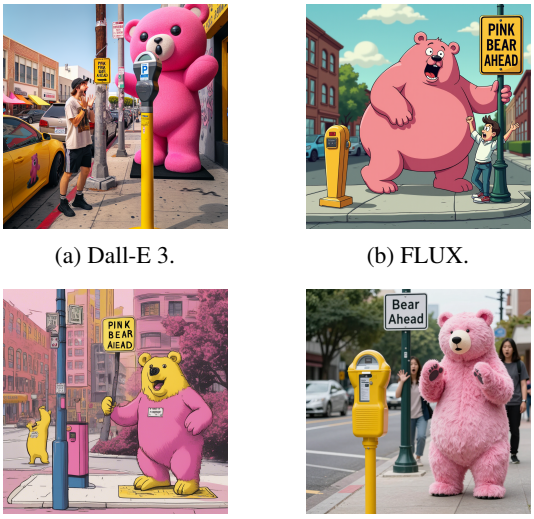
prompt for reference, for different models. All generations were conducted on a shared compute infrastructure comprising one NVIDIA RTX 6000 Ada Generation GPU with 49 GB of VRAM and two NVIDIA RTX 5000 GPUs with 32 GB of VRAM each. The system possessed an Intel Xeon w5-3435X (4.70 Ghz) CPU featuring 32 threads.

Based on our resource constraints for model loading and inference, for proprietary models, we selected DALL-E 3 (Betker et al., 2023) from OpenAI, partly to assess if a model from the same developer as our prompt generation model would perform better, and Gemini 3.1-Flash-preview (Google DeepMind, 2026) for which we acquired a license. For open-source alternatives, we included Z-Image Turbo (Cai et al., 2025), Stable Diffusion XL (Podell et al., 2024), FLUX 1.0-dev (Black Forest Labs, 2024) and Stable Cascade based on the Würstchen architecture (Pernias et al., 2024), to ensure a diverse set of models. Although we originally intended to use Qwen-Image (Wu et al., 2025) and Hunyuan-Image-3.0 (Cao et al., 2025) from Tencent, they were ultimately excluded as we were unable to load their full versions and wished to avoid comparisons with unofficial or distilled variants. To minimize developer bias, since we already included two models from the same distribution, we excluded GLM-Image which shares developers with Z-Image Turbo (Alibaba), and DeepFloyd (Saharia et al., 2022) (Stability AI, Stable Cascade).

We used the optimal inference parameters from the documentation of each model, as detailed in App.B. Following this approach, we generated >5,500 images for the 6 models. When a prompt exceeded the maximum sequence length of a model, we used SD-Embed (Zhu), 2024) (Apache License 2.0) to encode the text as learned embeddings.

3.3.3 Annotation process

To ensure diverse perspectives in the annotation process, we engaged 14 annotators from varied cultural and educational backgrounds (App.D for more details). Each distribution was annotated by three individuals, and annotations were merged upon manual verification, except for the less-objective *Cohesiveness* and *Emotion*, for which we solve disagreements based on majority votes. Our first merging method focused on annotator agreement, however, we found that some specific instances were too much confusing for annotators, such as those represented in App.G where humans are mixed with cats, which led to a global confusion. We proceeded to manual validation to ensure fair results.



(a) Dall-E 3. (b) FLUX.

(c) Stable Diffusion. (d) Z-Image Turbo.

Figure 3: Examples generated by various models.

Files were named according to the convention: `[model_name]_[skills]_[level]_[prompt number]_[robust/nothing]`. For example, the image in Fig.3d is named *z_image_turbo_2_easy_001_0.png*, indicating an easy-level prompt generated by Z-Image-Turbo for the *Color* + *Emotion* + *Text* skill combination. For the annotation process, each skill combination was assigned a specific code (e.g., 12 for "color+size"). Annotators first identified the model name to locate the corresponding folder (e.g., *flux*) and then applied the criteria outlined in Sec.3, along with column-specific rules. Annotation rules varied by skill. For *Counting*, values were recorded in the format `Instances_generated,Instances_required` for each object, separated by semicolons, allowing us to track over- or under-generation. For *Text*, annotators transcribed all visible text from signs or boards, separated by semicolons, with each sign's text on a new line. This approach helped capture instances where the model generated additional or incorrect text. For other skills, annotators recorded success rates, except for *Cohesiveness*, which was strictly binary (TRUE or FALSE). This process resulted in **+8,000 validated annotated raw rules**.

4 Experiments results

4.1 Metrics

In addition to the success rates presented in Sec.3, we utilize the Word Error Rate (WER) (Morris et al., 2004) to measure the discrepancy between

the generated text and the target text specified in the prompt. A lower WER indicates higher accuracy, with a value of 0 representing an exact match between the generated and reference texts. We opted for WER over embedding-based metrics such as cosine similarity because our objective was to assess the exact lexical match between the generated and target texts, rather than their semantic similarity.

4.2 Evaluation settings

As our benchmark comprises approximately 114 possible skill combinations, we focus on the *All Skills* setting in the remainder of this paper. This configuration represents the highest compositional load, requiring models to satisfy all evaluation dimensions simultaneously. Unless otherwise specified, the reported results are based on manual annotations and jointly assess Counting, Spatial Relations, Size Relations, Color Attribution, Emotion Attribution, Text Rendering, and Cohesiveness.

4.3 Overall results

The results presented in Tab.1 indicate that the Gemini-Flash-3.1 model achieved superior overall performance across most metrics (4/7) when considering all difficulty levels, with the exception of the Counting skill. However, a closer examination of *Cohesiveness* reveals that it failed to generate coherent images in 32% of the cases. WER values greater than 1 suggest that most models tended to generate either significantly longer sequences or additional sentences compared to the required text.

4.4 Analysis by level and number of skills

Analyzing performance across difficulty levels further elucidates the conditions under which models tend to fail, particularly as the required number of generated instances increases. Fig.4a, which aggregates results across all evaluated models, shows a highly non-uniform degradation across skills. The most pronounced decline from *easy* to *hard* occurs for the *Text* skill, with a 243% increase in error rate, indicating a substantial rise in hallucinated textual content. This is followed by *Counting*, *Cohesiveness*, and *Emotion*, which exhibit accuracy decreases of 71%, 60%, and 47%, respectively. In contrast, *Color* remains the most robust skill, with a comparatively modest drop of 19%.

Further insight is obtained by analyzing skill performance as a function of the number of combined skills. As shown in Fig.4d, even at a fixed difficulty level (here, *easy* level), accuracy typically decreases as the number of required skill combinations increases. This pattern reinforces our core hypothesis that generation quality is jointly impacted by both the number of instances to generate and the compositional load induced by multiple simultaneous constraints. While Fig.4d focuses on a single difficulty level, the same degradation is observed across all difficulty levels.

4.5 Prompt length vs Compositional load

While the dataset statistics (App. C) indicate that higher difficulty levels are generally associated with longer prompts and increased linguistic complexity (e.g., higher adjective density and greater use of prepositions), these aggregate trends remain incomplete. To control for potential confounding effects of prompt length, we isolate a subset of prompts with comparable lengths across all difficulty levels. Specifically, we estimate the probability distributions of prompt lengths for each level, identify the region with maximal overlap across distributions, and select the interval corresponding to the highest shared density. This procedure yields a length range of 50–77 tokens, comprising approximately 100 prompts per level. Fig.4c reports results restricted to this interval, aggregated across all models, on >2K annotated rules. The same trend persists: performance across most skills consistently degrades as difficulty increases. Importantly, we show in Tab. 2 that within this controlled subset, standard measures of linguistic complexity do not vary significantly with difficulty level.

Moreover, the standardized multivariate regression analysis shown in Tab. 3 indicates that compositional load accounts for a substantially larger share of performance variation than prompt length. Across all structured skills, compositional load yields larger standardized coefficients and lower p-values, while its confidence intervals remain consistently below zero, providing strong evidence of a negative association between increasing compositional complexity and model accuracy. By comparison, the effect of prompt length is generally smaller and less stable. Furthermore, VIF values below 5 for all predictors suggest that multicollinearity is limited, supporting the interpretation that compositional load contributes explanatory power beyond that captured by prompt length alone.

4.6 Analysis of TIIF-Bench

TIIF-Bench (Wei et al., 2025) reports a correlation between prompt length and the average skill

Skill	Z-Image-Turbo	FLUX 1.0	Dall-E 3	**Gemini-Flash**	SDXL	SC
↑ *Counting*	0.55	0.56	0.42	**0.70**	0.17	0.13
↑ *Spatial*	**0.79**	0.69	0.59	0.73	0.34	<0.01
↑ *Size*	0.84	0.62	0.70	**0.85**	0.15	0.20
↑ *Emotion*	0.80(0.94)	0.67(0.92)	0.28(1.00)	**0.98**(0.98)	0.09(1.00)	0.10(1.00)
↑ *Color*	**0.98**	0.88	0.69	0.96	0.16	0.36
↑ *Cohes.*	0.50(0.90)	0.76(0.90)	**0.99**(0.99)	0.68(0.98)	0.26(1.0)	0.35(1.0)
↓ *Text(WER)*	0.49	2.46	4.25	**0.29**	5.55	4.31

Table 1: Results for the "all skills" test, on manual annotation. FLUX 1.0=FLUX 1.0-dev, Gemini-Flash=Gemini-Flash-3.1-preview, SDXL=Stable Diffusion XL, SC=Stable Cascade. For the interpretative skills Emotion and Cohesiveness, values in parentheses report pairwise Cohen's κ, averaged across annotator pairs. Detailed results are available in App.E. For most categories, a higher score means a better evaluation, except for WER.

Criteria (avg)	**Easy**	**Medium**	**Hard**
Length	65.37	63.15	**67.31**
Entities	18.37	17.02	**18.94**
Relations	4.78	4.66	**4.92**
Parse depth	2.37	**2.47**	2.25
Prepositions	6.85	7.04	**7.22**

Table 2: Statistics on prompts of similar lengths, obtained using NLTK tokenizer. Entities=Nouns, Relations=Number of connectors (with, on, under...), Parse depth=Number of clauses (which, that...) and sentences.

accuracy. However, a closer examination of their prompt construction reveals an intriguing pattern.

Consider the example in Fig. 5. By analogy with our prompt construction, the shorter prompt can be interpreted as involving two instances (*pig*,*cup*) and a single *Spatial* rule (*to the left of*). In contrast, the longer variant includes the original compositional load and increases it with additional descriptive details such as texture attributes. In our benchmark, while such descriptive elements may be included, they are not systematically scaled with difficulty. Instead, variations in prompt length arise mainly from the controlled factors. Our initial analysis compared the lexical content of the long and short prompt variants and showed that the additional content was predominantly related to scene atmosphere, texture descriptions, and contextual details. We subsequently formalized these recurring additions as additional "rules" and re-analyzed the benchmark from that perspective. Under this procedure, more than 90% of the analyzed prompt pairs exhibited an increase in compositional content beyond simple length expansion.

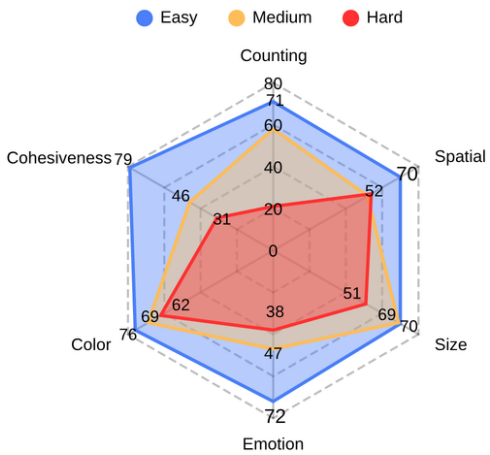
Although prompt length can influence model performance by increasing the burden on the text encoder to faithfully capture constraints, we argue that disentangling length from compositional structure is critical for a more interpretable assessment.

(Short): A cup is positioned to the left of a pig.
(Long): Positioned tranquilly to the left of the pig, which stands as a silent and innocent observer, the unassuming cup, a simple vessel of ceramic or maybe porcelain, rests quietly, its presence understated yet somehow integral to the quietude of the scene, where each object seems to hold its breath in the gentle stillness that pervades the atmosphere.

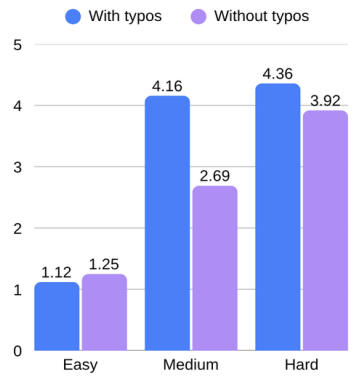
Figure 5: Example of two versions of a prompt, from TIIF-Bench(Wei et al., 2025). Parts of the text that are highlighted represent potential "rules" or instances.

4.7 T2I models struggle at typos

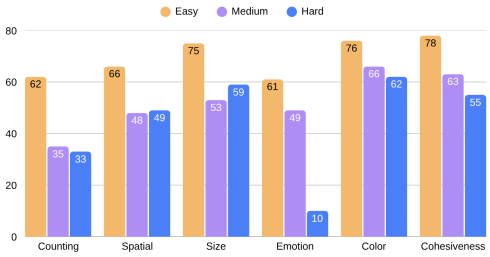
Fig.4b compares text rendering performance for prompts with and without injected typos. Across most difficulty levels, they exhibit substantially larger deviations in generated text. A closer inspection of failure cases under typo-conditioned



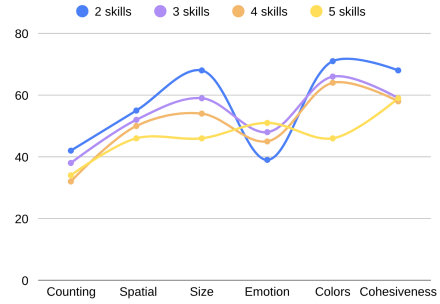
(a) Skills by level.



(b) WER with and without typos.



(c) Results on similar prompt lengths.



(d) Analysis by number of skill combinations (rules).

Figure 4: Breakdown of evaluation metrics (avg) on multiple axis. More complete results available in App. F.

prompting shows that, beyond hallucinations, models often either introduce alternative spelling errors rather than the specified ones (30% of cases).

5 Discussion

5.1 The importance of human annotations

We developed an automated pipeline to streamline the annotation process. **Object detection, monocular depth estimation and and instance segmentation** using YOLO-26 (Jocher et al., 2026) (AGPL-3.0), extracts bounding boxes and class labels, depth maps and detection masks, enables automated evaluation of the *Counting, Spatial* and *Size* skills. *Color, Emotion* and *Cohesiveness* rely on a Qwen3-VL (Yang et al., 2025) inference (Apache License 2.0).

However, the pipeline struggled significantly with *Cohesiveness*, for which (VQA) model alignment with human judgment rarely exceeded 70%. We explored various methods to improve those assessments, drawing on recent MLM-as-judge studies (Chen et al., 2024), including instance-level evaluation and prompt reformulation. However, results remained unsatisfactory. Traditional neural networks based frameworks such as DeepFace (Serengil and Ozpinar, 2020) also failed to deliver significant improvements. This further reinforces the importance of human oversight for interpretative dimensions.

5.2 The state of current evaluation

Our findings, especially when reviewing the state-of-the-art, raised an intriguing question: ***Are our current evaluation methods too "static"?*** To the best of our knowledge, most existing benchmarks (including ours) rely on a fixed set of skills or predefined combinations thereof. A promising direction for future evaluation lies in the development of *modular* benchmarks, in which skills can be added or removed on demand and corresponding evaluation prompts can be generated accordingly. Such adaptability would enable continuous refinement of evaluation protocols as model capabilities evolve.

6 Conclusion

This work introduced IMAG-EVAL, a controlled and interpretable evaluation framework aimed at diagnosing the instruction-following capabilities of Text-to-Image (T2I) models. Motivated by the limitations of existing benchmarks, we focused on dis-

Skill	Coefficients	P-values	Confidence intervals	VIF (L,CL)
Counting	(L) $+0.0216$ **(CL) -0.3634**	(L) 0.61 **(CL) 9.8×10^{-17}**	(L) $[-0.06; 0.10]$ (CL) $[-0.44; -0.27]$	3.52
Spatial	(L) $+0.0341$ **(CL) -0.1910**	(L) 0.29 **(CL) 4.89×10^{-9}**	(L) $[-0.029; 0.09]$ (CL) $[-0.25; -0.12]$	2.83
Size	(L) $+0.0418$ **(CL) -0.1883**	(L) 0.24 **(CL) 2.13×10^{-7}**	(L) $[-0.02; 0.11]$ (CL) $[-0.25; -0.11]$	3.15
Emotion	**(L) -0.1203** (CL) -0.0130	**(L) 0.001** (CL) 0.73	(L) $[-0.19; -0.04]$ (CL) $[-0.08; 0.06]$	3.03
Colors	(L) -0.0074 **(CL) -0.1296**	(L) 0.82 **(CL) 0.0001**	(L) $[-0.07; 0.05]$ (CL) $[-0.19; -0.06]$	3.50
Text (WER)	(L) $+0.0910$ **(CL) $+0.1548$**	(L) 0.017 **(CL) 0.000059**	(L) $[0.01; 0.16]$ (CL) $[0.07; 0.23]$	3.07
Cohesiveness	**(L) $+0.0732$** (CL) $+0.0159$	**(L) 0.008** (CL) 0.56	(L) $[0.01; 0.12]$ (CL) $[-0.03; 0.07]$	2.88

Table 3: Results of the multivariate regression and correlation analyses: skill accuracy vs (prompt length and compositional load). Abbreviations: CL=Compositional Load, L=Length, VIF = Variance Inflation Factor. All the coefficients were standardized prior to model fitting. Across most structured skills (Counting, Spatial, Size, Color, and Text), compositional load has a strong and statistically significant effect, whereas prompt length exhibits a weaker effect. In contrast, for interpretative skills (Emotion and Cohesiveness), this relationship is less pronounced.

entangling surface linguistic complexity from compositional difficulty, and on capturing failure modes that are often overlooked in standard skill-based evaluations, such as global incoherence and incomplete object generation. We further introduced a benchmark dataset comprising 1,140 prompts and +8,000 combined rules, and validated our approach through extensive experiments involving more than 6,000 generated images, 14 annotators, and +8,000 annotated rules. These results demonstrate both the feasibility and the diagnostic value of controlled, compositional evaluation protocols.

Our empirical findings highlight that model performance is primarily driven by compositional load (the number of grounded rules and their binding to instances) rather than by surface-level linguistic properties such as prompt length, for structured skills. More broadly, they underscore the limitations of evaluation practices that rely on single-axis proxies for difficulty, and motivate the need for benchmarks that provide a more structured decomposition of multimodal reasoning.

Looking forward, an important direction is the design of flexible and modular benchmarks for multimodal evaluation. While our current framework already enables controlled variations (e.g., adjusting instance counts or regenerating prompts with fixed constraints), a natural extension is to support fully modular skill composition, where evaluation dimensions can be dynamically added, removed, or recombined. Such flexibility would enable more systematic stress-testing of model capabilities and foster the development of more robust and interpretable multimodal systems.

Acknowledgments

This work was funded by Talan France through its Research and Innovation Center. We also thank all annotators for their invaluable contributions to data collection and quality assurance, in particular Mariem AMMAR, Dodji Idelphonse DECADJEVI and Samia TEKAL, for their exceptional commitment to the project.

Limitations

While our evaluation framework demonstrates promising results, it is currently limited to prompts constructed from COCO object categories and attributes emotions exclusively to human instances. Nevertheless, COCO covers a broad spectrum of common objects, enabling the use of a wide range of off-the-shelf object detectors for automated evaluation, as most contemporary detectors are trained

and benchmarked on COCO.

Moreover, all images in this study were generated using a single random seed (42). Evaluating multiple seeds would have required generating and annotating a substantially larger number of samples, resulting in annotation costs beyond the resources available for the present work. Nonetheless, robustness across random initializations is an important consideration, and future versions of the leaderboard will include multi-seed evaluations together with standard deviation estimates.

A further limitation is that our method evaluates single-prompt instructions rather than multi-prompt, incremental instruction sequences. Although some models may perform better with step-by-step guidance, we lack a clear methodology for determining the optimal order of rule presentation and its impact across models, due to the potential number of combinations of orders and skills.

Finally, our evaluation may be subject to two sources of bias.

(1) Language bias. All prompts in Imag-Eval are formulated in English. Although all evaluated models officially support English-language inputs, our findings may not directly generalize to languages with substantially different typological, morphological, or syntactic properties. Extending the benchmark to multilingual settings is therefore an important direction for future work.

(2) Model-identity bias. The annotation protocol did not fully blind annotators to model identity, as generated images were organized using filenames and folders that included model names. This design choice simplified annotation management and downstream analysis, but may have introduced bias in subjective judgments. We partially mitigate this concern through multiple annotators, majority voting for interpretative skills, and manual validation of ambiguous cases. Moreover, most annotators had limited familiarity with the evaluated T2I systems beyond widely known commercial models such as Gemini-Flash-3.1. Nevertheless, future versions of the benchmark should adopt fully anonymized filenames and randomized annotation interfaces.

References

Arash Afkanpour, Omkar Dige, Fatemeh Tavakoli, Negin Baghbanzadeh, Farnaz Kohankhaki, and Elham Dolatabadi. 2025. Automated Capability Evaluation of Foundation Models. *arXiv preprint arXiv:2505.17228*.

Nefeli Andreou, Varsha Vivek, Ying Wang, Alex Vorobiov, Tiffany Deng, Raja Bala, Larry Davis, and Betty Mohler Tesch. 2024. BodyMetric: Evaluating the Realism of Human Bodies in Text-to-Image Generation. *arXiv preprint arXiv:2412.04086*.

Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. 2023. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20041–20053.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *OpenAI Technical Report*.

Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Black Forest Labs. 2024. Flux.1 [dev] model card. Hugging Face Model Card.

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

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, and others. 2025. HunyuanImage 3.0 Technical Report. *arXiv preprint arXiv:2509.23951*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. MLLM-as-a-Judge: assessing multimodal LLM-as-a-Judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models. In *ICCV*.

Ciprian A Corneanu, Qianli Feng, and Aleix M Martinez. 2025. Structured human assessment of text-to-image generative models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4481–4490. IEEE.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Google DeepMind. 2026. Gemini 3.1 flash image: Model card. Google DeepMind.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, and Shengfeng Ye. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

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

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

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2025. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579.

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

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. 2026. Ultralytics yolo26: Unified real-time end-to-end vision models. *Preprint*, arXiv:2606.03748.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Zachary C Lipton. 2018. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Proc. Interspeech 2004*, pages 2765–2768.

Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. 2024. W"urstchen: An Efficient Architecture for Large-Scale Text-to-Image Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

Vitali Petsiuk, Alexander E Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A Plummer, Ori Kerret, and others. 2022. Human evaluation of text-to-image models on a multi-task benchmark. *arXiv preprint arXiv:2211.12112*.

Dustin Podell, Zion English, and Kyle Lacey. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *International Conference on Learning Representations*, volume 2024, pages 1862–1874.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj"orn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Sefik Ilkin Serengil and Alper Ozpinar. 2020. Lightface: A hybrid deep face recognition framework. In *2020 innovations in intelligent systems and applications conference (ASYU)*, pages 1–5. IEEE.

Aaditya Singh, Adam Fry, and others. 2025. OpenAI GPT-5 System Card. _eprint: 2601.03267.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. 2025. TIIF-Bench: How Does Your T2I Model Follow Your Instructions? *arXiv preprint arXiv:2506.02161*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, and Kaiyuan Gao. 2025. Qwen-Image Technical Report. _eprint: 2508.02324.

An Yang, Anfeng Li, and Baosong Yang. 2025. Qwen3 Technical Report. _eprint: 2505.09388.

Yue Yu, Ding Li, Benyuan Li, and Nengli Li. 2024. Multi-style image generation based on semantic image. *The Visual Computer*, 40(5):3411–3426.

Shudong Zhu(Andrew Zhu). 2024. Long prompt weighted stable diffusion embedding. https://github.com/xhinker/sd_embed.

A Modifying levels and instance count

As discussed earlier, our JSON-based architecture is designed to flexibly redefine both the granularity of difficulty levels and the number of instances to generate at each level. As illustrated in Fig.6, adding any level to the list of predefined levels (top panel) and any cases within the cases (bottom panel) allow for re-defining levels. Re-running the main prompt-generation script after such changes automatically produces a new set of prompts reflecting the updated parameters.

In addition, the classes that govern prompt generation within the scripting interface (`llm_interfaces`) are model-agnostic. Users may specify their own API keys and model parameters via a `.env` file, provided the required configuration fields are defined. As a result, the framework is not restricted to GPT-5 or OpenAI models, and can be readily extended to evaluate a wide range of proprietary or open-source text-to-image systems.

B Image generation hyperparameters

Tab.4 showcases hyperparameters used during our image generation process. All parameters used, except random seeds, were set according to the official report of each of the models.

C Dataset statistics

Our dataset includes (combined) 2,228 *Counting* rules, 1,660 *Color*, 2,199 *Spatial*, 372 *Emotion*, 2,197 *Size*, and 186 *Text*. Cohesiveness is not instantiated as an explicit rule; instead, it is enforced globally throughout the evaluation protocol. The distributions of all dataset elements are reported in Fig.7.

D Annotator Recruitment and Instructions

The annotators were recruited on a voluntary basis and selected to ensure diversity in educational background, gender, and cultural perspective. All annotators received a README, outlining the study goals, the role of their annotations in the evaluation pipeline, and the procedures for anonymization. Annotators were informed that parts of the annotations may be included in the submission for transparency, while strictly preserving anonymity throughout the review and publication process. The recruitment procedures were conducted according to the guidelines of our institutional ethics board.

The README included comprehensive annotation guidelines, accompanied by examples for each skill. Fig.9 illustrates an example of these instructions. While most evaluation criteria were designed to be straightforward (e.g, number of instances), the *Cohesiveness* skill required more interpretative judgment. To ensure consistency, we provided illustrative examples covering a range of failure modes, including incomplete objects (e.g., an airplane without wings), anatomical inconsistencies (e.g., missing/extra body parts), distortions (e.g., distorted facial features), and implausible configurations (e.g., unsupported floating objects).

We also documented representative edge cases, including ambiguous generations, and provided explicit guidelines on how to handle such cases during annotation. Fig.8 shows an example of the annotation interface distributed to annotators.

E Overall results for all the models

Results for all skill combinations are reported in Tab. 5. We exclude GEMINI-3.1-FLASH from the

```python
class MetaPromptGeneration:
    def __init__(self, metadata):
        self.levels = ["easy", "medium", "hard"]
        self.data_type = "coco"
        self.metadata = metadata
        self.scene = {}

    def _setup_scene(self, skills, level):
        scene = {}
        if Skills.EMOTION in skills:
            scene = self.apply_emotion(level, scene)
            if skills==[Skills.EMOTION]:
                self.scene = scene
                return scene

    def apply_counting(self, level, scene):
        low, high = 0, 0
        k = len(scene['objects'])
        match level:
            case "easy": low = 1; high = 2
            case "medium": low = 2; high = 3
            case "hard": low = 4; high = 5
```

Figure 6: Configuration parameters controlling difficulty granularity and the number of generated instances.

overall analysis because a subset of images could not be generated due to content filtering and API budget/limit constraints.

F Overall results and impact of Compositional load

Results from prior sections indicate that, even under fixed difficulty (i.e., a constant number of instances), performance consistently degrades as the number of required skills increases. Fig. 10 extends this observation across all difficulty levels and most skill categories.

Fig. 11 further analyzes this effect through Pearson correlation heatmaps, comparing prompt length (left) and compositional load (right) against skill-specific accuracies. Across the majority of skills, compositional load exhibits substantially stronger correlations with performance than prompt length. The only exceptions are *Cohesiveness* and *Emotion*, for which correlation differences remain marginal.

Moreover, our regression analysis illustrated in Tab.3 display the same trend. We report multivariate linear regression coefficients, along with standard errors, p-values, and 95% confidence intervals, computed using ordinary least squares. All variables are standardized prior to regression, enabling direct comparison of coefficient magnitudes. Across most skills, compositional load has a strong and statistically significant negative effect, whereas prompt length exhibits a small and non-significant effect, with the only exceptions being the most interpretative skills, *Cohesiveness* and *Emotion*. Confidence intervals also suggest that compositional load has a more consistent effect on a decrease in accuracy.

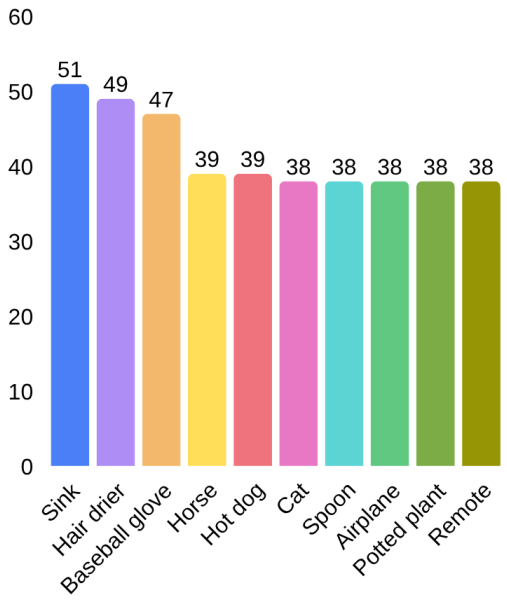
Taken together, these findings suggest that performance is more strongly governed by compositional load than by prompt length alone for structured skills, highlighting the importance of evaluating models along multiple, complementary axes of difficulty.

G Confusing cases

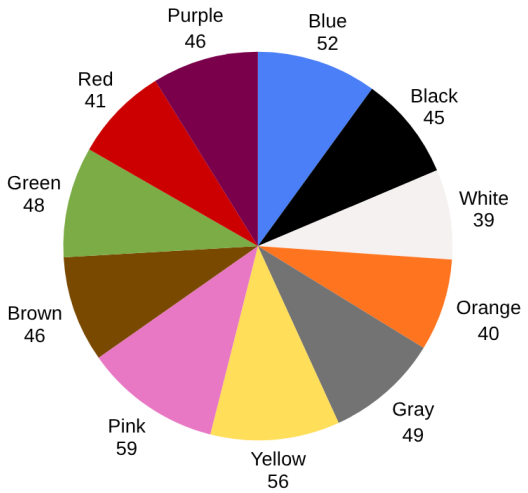
Examples of confusing cases that lead us to manual validation for all annotations are available in Fig.12. Some include generations where instances could not be determined (such as in Analysis Fig.12a) or the model hallucinated too much to take some results into account (Fig.12b).

H Use of AI assistants

The use of LLMs was limited to rephrasing and formatting assistance and did not affect the scientific method, experimental design, implementation,



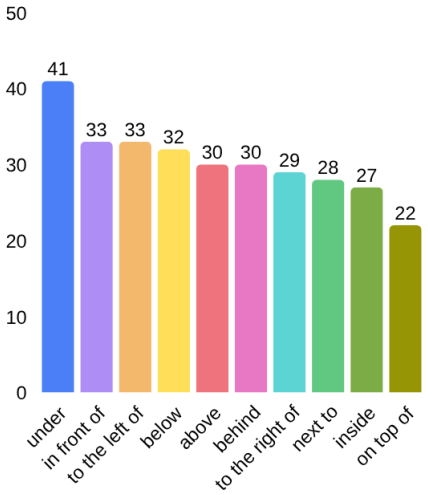
(a) Top 10 objects within the dataset.



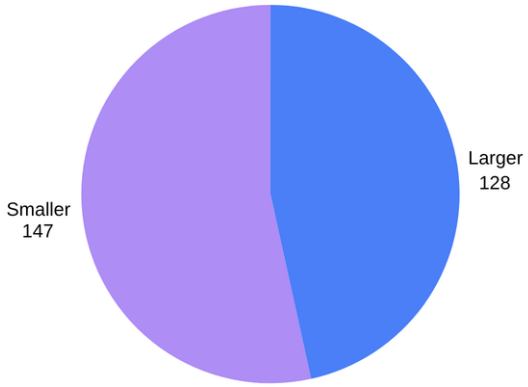
(b) Color distribution in the dataset.

Criteria (avg)	Easy	Medium	Hard
Length	69.35	90.26	**109.21**
Entities	19.80	24.97	**30.67**
Relations	5.07	7.08	**8.61**
Parse depth	2.38	3.46	**3.66**
Adjectives density	**0.16**	0.13	0.13
Prepositions	7.31	9.95	**11.53**

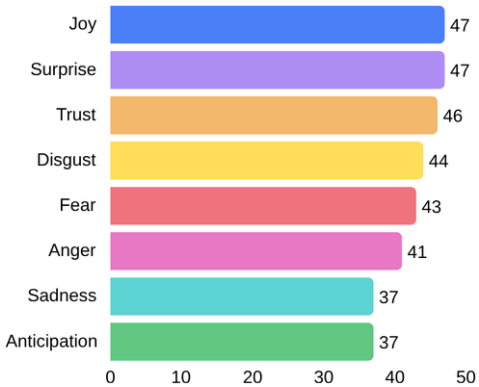
(c) Textual statistics. Reported values are averages. Entities=Number of nouns, Relations=Number of connectors and prepositions, Parse depth=Number of clauses and sentences, Adjectives density=(Number of adjectives/Total length), Prepositions=Number of prepositions. All English prompts.



(d) Frequency of spatial relationships.



(e) Size relationships.



(f) Distribution of emotions.

Figure 7: Our dataset (JSON prompt collections), annotation guidelines, and associated research assets will be released under the Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) license, while the accompanying source code will be released under the MIT License. These resources are intended solely for research, educational, and non-commercial R&D purposes. Textual statistics are extracted using the default tokenizer of NLTK (Bird and Loper, 2004). We will release validated annotation samples; generated images will be released only when permitted by the corresponding model licenses and API terms.

Model (license)	Method	Parameters	Values
All local models	-	Dimensions (w x h) Random seed	1024x1024 42
Z-Image-Turbo (APL 2.0)	Local	Guidance scale Inference steps	0 9
Stable Diffusion XL (CreativeML)	Local	Guidance scale Inference steps	5 50
FLUX 1.0-dev (Non-Commercial License)	Local	Inference steps Config scale Sample method	50 1 Euler
Stable Cascade (MIT license)	Local	Guidance scale Inference steps	3 30
Dall-E 3 (proprietary) Gemini-3.1-Flash-preview (proprietary)	API	Temperature	1

Table 4: Parameters for image generation, listed for reproducibility purposes. APL=Apache License. All models used were consistent with their intended use.

Skill	Z-Image-Turbo	FLUX 1.0	Dall-E 3	SDXL	SC
$\uparrow Counting$	**0.61**	0.57	0.41	0.27	0.13
$\uparrow Spatial$	**0.82**	0.71	0.52	0.28	<0.01
$\uparrow Size$	**0.83**	0.77	0.72	0.32	0.20
$\uparrow Emotion$	**0.94**	0.45	0.44	0.21	0.10
$\uparrow Color$	**0.92**	0.86	0.75	0.52	0.35
$\uparrow Cohesiveness$	**0.90**	**0.90**	0.71	0.11	0.35
$\downarrow Text(WER)$	**0.61**	1.73	4.21	4.92	4.31

Table 5: Aggregate across all potential skill combinations, from 2 to 6-skill combinations. FLUX 1.0 = FLUX 1.0-dev, SDXL = Stable Diffusion XL, SC = Stable Cascade. Z-Image-Turbo outperforms other evaluated models in most metrics. Gemini was excluded due to content filter and budget issues leading to missing images.

evaluation or originality of the research.

Comment remplir le fichier ? (Ex : stable_diffusion_0_easy_001_0.png)

1. **Retrouver l'image.** Ici, l'on remarque que le modèle est **stable diffusion**, il faudra donc se rendre dans le dossier prévu ici : stable_diffusion.
2. Vérifier le code des skills. Ici, l'on remarque que **le skill code est 0**. On vérifie à quoi cela correspond dans le fichier prévu. 0 correspond à "all_skills", donc l'on cherche à tester toutes les compétences (donc toutes les cases).
3. Lire le prompt associé dans votre fichier d'annotations (exemple : ce fichier). Ici, il est le suivant : "Photo realistic street scene with **two small blue fire hydrants** and **one large purple apple: two miniature blue fire hydrants (clearly smaller than the fruit) stand on a sidewalk** with **a single oversized purple apple positioned directly beneath and centered between them. A joyful person nearby, smiling broadly with arms raised, celebrates the whimsical sight. A small sign next to the apple reads Purple Apple.** Soft daylight, shallow depth of field, realistic details."
4. Annoter les cases :
   a. **Counting:** On vérifie les objets demandés au modèle. 4 *fire hydrants* demandés, 1 généré = 1,4. 1 *apple* demandé, 1 générée = 1,1. 1 *person*

Figure 8: A screenshot of the file containing instructions disclosed to the annotators. We included relevant image examples and how they will annotate relevant fields on the linked spreadsheet.

Image	Prompt	Counting	Colors	Spatial relationships	Size	Emotion	Text	Cohesiver
z_image_turbo_0_easy_001_0.png	Photorealistic street scene with two small blue fire hydrants and one large purple apple: two miniature blue fire hydrants (clearly smaller than the fruit) stand on a sidewalk with a single oversized purple apple positioned directly beneath and centered between them. A joyful person nearby, smiling broadly with arms raised, celebrates the whimsical sight. A small sign next to the apple reads Purple Apple. Soft daylight, shallow depth of field, realistic details.	2,2;1,1;0,2;0,1;1,1;1,1	1.00	1.00	1.00	1.00	Purple Apple	Vrai
z_image_turbo_0_easy_001_0_robust.png	A single large yellow airplane sits directly beside two small black sinks, the airplane noticeably larger than both sinks; a person nearby shows clear anger, scowling and pointing at the airplane. On a small sign above the sinks is visible text that reads "Yellow airplane" and no other words.	1,1;2,2;1,1;1,1	1.00	1.00	1.00	1.00	Yellow airplane	Vrai
z_image_turbo_0_easy_001_1.png	A whimsical outdoor scene with two small blue fire hydrants positioned above and to either side, and one oversized purple apple sitting directly underneath them so the apple is noticeably larger than the hydrants. A small wooden sign planted in front of the apple clearly reads "Purple apple". Nearby a joyful person beams with delight, smiling broadly and raising their arms in celebration.	2,2;1,1;1,1;1,1	1.00	1.00	0.50	1.00	Purple Apple	Vrai
z_image_turbo_0_easy_001_1_robust.png	A large lone yellow airplaine is positioned right beside two small black sinks, the airplaine clearly much larger than the sinks; a nearby persson is visibly angry, scowling and pointing at the plane. Above the sinks a small sign displays the words "Yellow airplane" and no other text.	1,1;1,1;1,1;2,1	1.00	1.00	1.00	1.00	Yellow aiplane	Vrai

Figure 9: A screenshot containing an example of spreadsheet given to each annotator. This corresponds to the z-image distribution.

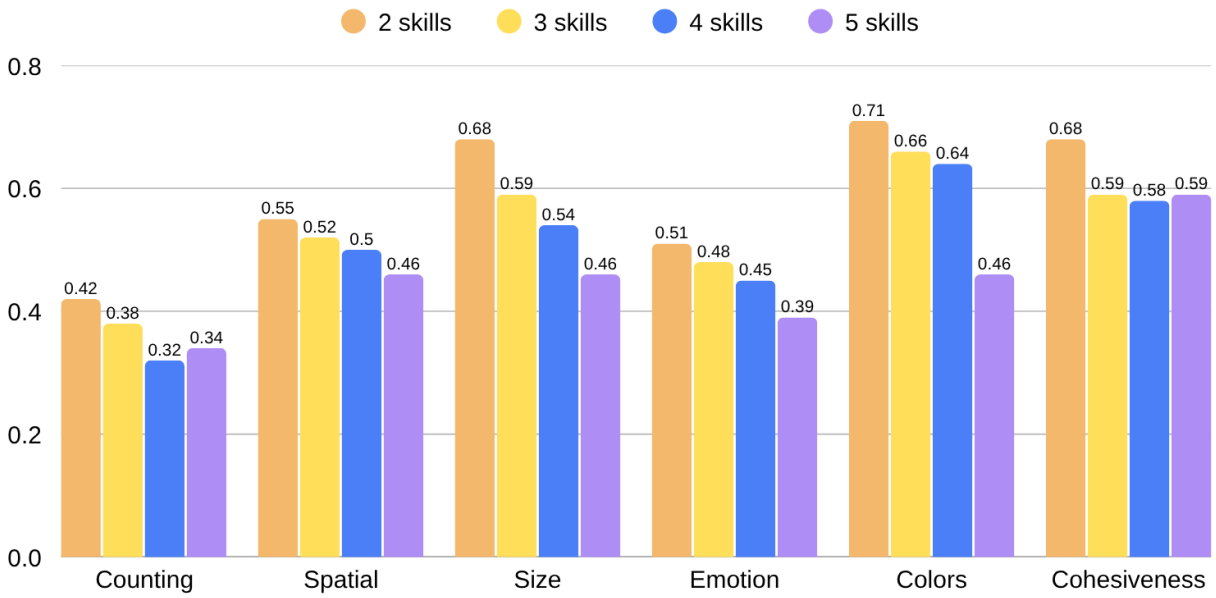


Figure 10: Skill accuracy as a function of the number of rules (skill combinations). A clear and consistent trend emerges: performance degrades substantially as the number of enforced textual rules increases, across nearly all skills. When considered jointly with our findings on the effect of the number of instances to generate, these results motivate the need for a two-factor analysis of task complexity.

Figure 11: Heatmap comparing prompt length and compositional load. Cells report Pearson correlation coefficients. Across most skills, increases in compositional load exhibit stronger negative correlations with performance than prompt length. Notable exceptions are *Emotion* and *Cohesiveness*, for which correlation differences remain marginal.

(a) Confusing case 1.

(b) Confusing case 2.

Figure 12: Examples of confusing cases that lead us to manual annotation validation. Humans are mixed with cats, which leads to various reported accuracies, depending on if they decided to consider the annotation as human, cats, or both. The second figure was generated by stable Diffusion XL with the same prompt as the first one, but it generated an incoherent scene.