T2I-FactualBench: Benchmarking the Factuality of Text-to-Image Models with Knowledge-Intensive Concepts

Ziwei Huang[1], **Wanggui He**[2], **Quanyu Long**[3], **Yandi Wang**[1], **Haoyuan Li**[2], **Zhelun Yu**[2], **Fangxun Shu**[2], **Long Chan**[2], **Hao Jiang**[2], **Fei Wu**[1], **Leilei Gan**[1*],

[1]Zhejiang University, [2]Alibaba Group, [3]Nanyang Technological University,
{ziweihuang, leileigan}@zju.edu.cn

Abstract

Most existing studies on evaluating text-to-image (T2I) models primarily focus on evaluating text-image alignment, image quality, and object composition capabilities, with comparatively fewer studies addressing the evaluation of the factuality of the synthesized images, particularly when the images involve knowledge-intensive concepts. In this work, we present T2I-FactualBench—the largest benchmark to date in terms of the number of concepts and prompts specifically designed to evaluate the factuality of knowledge-intensive concept generation. T2I-FactualBench consists of a three-tiered knowledge-intensive text-to-image generation framework, ranging from the basic memorization of individual knowledge concepts to the more complex composition of multiple knowledge concepts. We further introduce a multi-round visual question answering (VQA)-based evaluation framework to assesses the factuality of three-tiered knowledge-intensive text-to-image generation tasks. Experiments on T2I-FactualBench indicate that current state-of-the-art (SOTA) T2I models still leave significant room for improvement. We release our datasets and code at https://github.com/Safeoffellow/T2I-FactualBench.

1 Introduction

In recent years, text-to-image (T2I) generation have made significant advancements in synthesizing high-fidelity and diverse style images from input textual descriptions (Rombach et al., 2022; Podell et al., 2024; Li et al., 2024c; Chen et al., 2024a; Sun et al., 2024b,a; Wang et al., 2024c; He et al., 2024). T2I models, represented by diffusion models (Rombach et al., 2022; Podell et al., 2024; Li et al., 2024c; She et al., 2025) and autoregressive models (Sun et al., 2024b,a; He et al., 2024; Wang et al., 2025b; Liu et al., 2025), have been applied to a wide range of scenarios, including e-commerce, art and games (Oppenlaender, 2022; Vashishtha et al., 2024; Vimpari et al., 2023).

A significant challenge accompanying the advancement of T2I generation lies in evaluating the generated images (Hartwig et al., 2024). Most existing efforts on this challenge primarily focus on evaluating text-image alignment (Hessel et al., 2021; Radford et al., 2021; Yu et al., 2022), image quality (Kirstain et al., 2023; Li et al., 2024b; Xu et al., 2024; Saharia et al., 2022; Wu et al., 2023) and object composition capability (Park et al., 2021; Saharia et al., 2022; Li et al., 2024a; Wu et al., 2024b; Huang et al., 2023) *inter alia*, using automated metrics such as Fréchet Inception Distance (FID) (Heusel et al., 2017), Inception Score (IS) (Salimans et al., 2016), and CLIPScore (Hessel et al., 2021). Recently, several efforts have been made to evaluate the reasoning capabilities of T2I models, as exemplified by Commonsense-T2I (Fu et al., 2024) and PhyBench (Meng et al., 2024).

However, despite the aforementioned efforts, a comprehensive benchmark for evaluating the factuality of T2I models in generating knowledge-intensive concepts and their compositions is still lacking. Knowledge-intensive concepts differ significantly from general concepts or objects because their visual features are often difficult—or even unnecessary—to explicitly describe in the input textual description. For example, as shown in Fig. 1, when given prompts with general concepts, the SOTA T2I model effectively generate images that fulfill the instructions. However, when presented with specific knowledge-intensive concepts, such as a *LV M43986 Cannes handbag*, the generated images often struggle to accurately represent the intended concepts. This characteristic sets the evaluation of such concept generation apart from traditional text-alignment evaluations.

To our best knowledge, the most closely related studies on the evaluation of knowledge-intensive

*Corresponding Author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27501–27524

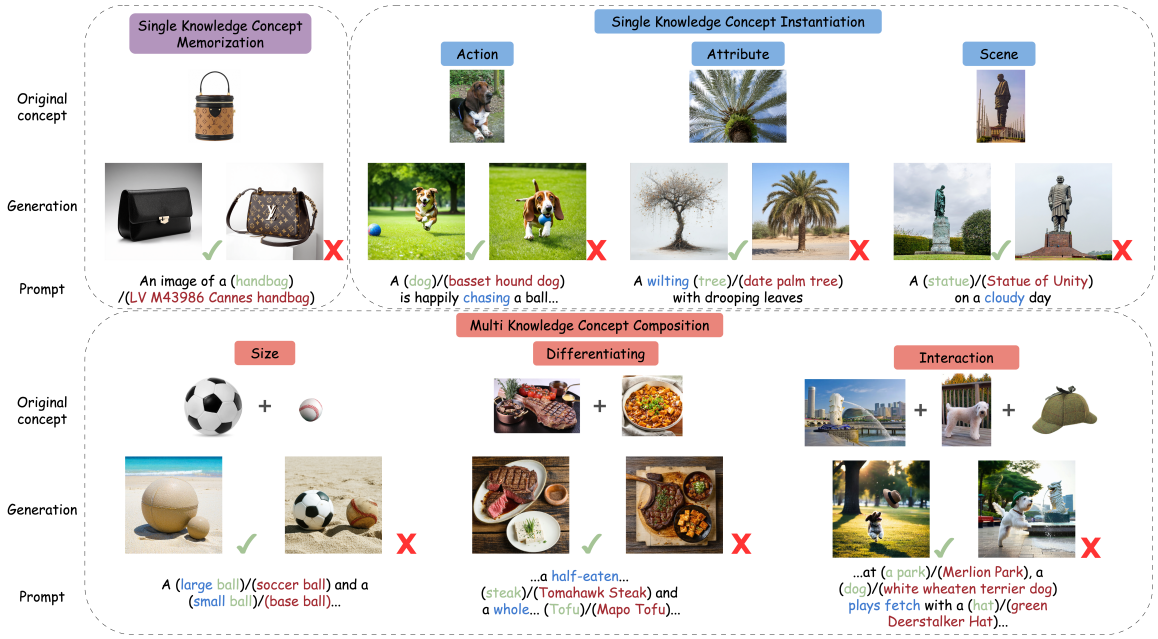

Figure 1: **General Concepts vs. Knowledge Concepts.** We use the SOTA T2I model Stable Diffusion 3.5(SD 3.5; (Esser et al., 2024)) as an example to illustrate the challenges posed by knowledge-intensive concepts versus general concepts. When given prompts with general concepts (indicated in green), SD 3.5 effectively generates images (left in Generation) that fulfill the instructions. However, when presented with specific knowledge concepts (indicated in red), SD 3.5 (right in Generation) often struggles to meet the requirements or accurately represents the intended concepts. This issue is particularly pronounced when the images are required to compose multiple knowledge concepts. The blue text in the prompts highlights the specific tasks to be achieved.

concept generation are as follows: HEIM (Lee et al., 2024) conducts a holistic evaluation of T2I models across 12 different aspects, such as alignment, quality and aesthetic, etc. Among them, the knowledge dimension evaluate whether the model have knowledge about the world or domains. However, HEIM evaluates only a very limited set of real-world entities and employs superficial CLIP-Score to assess the factuality of entities. We also note the concurrent work KITTEN (Huang et al., 2024) explores the knowledge-intensive evaluation of image generation for real-world visual entities. However, KITTEN employs four pre-defined templates to generate input textual descriptions, restricting its flexibility to provide a comprehensive evaluation of concept under different scenarios.

In this work, we present T2I-FactualBench, the largest benchmark to date in terms of the number of concepts and prompts designed to evaluate the factuality of T2I models when generating images that involves knowledge-intensive concepts. The construction of T2I-FactualBench begins with collecting **a set of Knowledge Concepts, which are defined as concepts with a limited number of hyponyms in the knowledge base**. Knowledge concepts are specifically designed to challenge T2I models by requiring them to precisely generate inherent visual details of each concept. Building upon the collection of knowledge concepts, we next propose a three-tiered knowledge-intensive text-to-image generation framework, spanning from the basic memorization of individual knowledge concepts to the more complex composition of multiple knowledge concepts.

To conduct an effective and efficient evaluation of existing T2I models' performance on the proposed T2I-FactualBench, we also introduce a multi-round visual question answering (VQA)-based evaluation framework aided by advanced multi-modal LLMs. This multi-round VQA evaluation framework firstly assesses the factuality of the generated concept with respect to the reference image, and then evaluates the completeness of concept instantiation under different conditions, lastly examines the factuality of multiple concept compositions under varying scenarios.

We conduct a comprehensive evaluation of the performance of seven closed- and open-source T2I models on the proposed T2I-FactualBench, such as Stable Diffusion models (Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024), Flux.1 (Andreas Blattmann, 2024) and DALLE-3 (Betker et al., 2023). Furthermore, we explore two approaches for injecting external knowledge into the

models to facilitate the generation of knowledge concepts. The first approache is *Visual-Knowledge Injection*, where images of knowledge concepts are provided as references to guide the model in image generation. The second is *Text-Knowledge Injection*, where we provide textual descriptions of the visual features of knowledge concepts as external knowledge. Experiments on T2I-FactualBench indicate that current state-of-the-art (SOTA) T2I models still leave significant room for improvement.

2 Related work

2.1 Text-to-Image Generation Evaluation

Current evaluation metrics for text-to-image generation focus on key aspects such as the fidelity of generated images, assessed using FID and IS (Heusel et al., 2017; Salimans et al., 2016), and perceptual similarity evaluated by CLIP-I and DINO Score (Caron et al., 2021). Additionally, image-text alignment is measured using metrics such as CLIP-T, CLIP Score, and BLIP Score (Radford et al., 2021; Hessel et al., 2021; Li et al., 2022). However, these metrics often fall short in capturing the intricate nuances of text-image alignment. With the advancement of large language models (LLMs) and large multimodal language models (MLLMs) (Chen et al., 2024b; Wang et al., 2024a; OpenAI, 2024; Xiao et al., 2024b; Zhao et al., 2024; Xiao et al., 2024a; Liu et al., 2024; Lin et al., 2024; Bi et al., 2025, 2024), some approaches use MLLMs to employ VQA evaluation (Hu et al., 2023; Lu et al., 2024b; Huang et al., 2023; Kirstain et al., 2023; Xu et al., 2024). However, the binary yes-or-no answer format proves insufficient for detailed evaluations. Human evaluations (Ku et al., 2023b; Lu et al., 2024a; Huang et al., 2024) provide critical insights but are limited by significant costs and time-intensive processe. Some studies (Ku et al., 2023a; Zhang et al., 2023a; Wu et al., 2024a; Peng et al., 2024) highlight GPT-4V's potential in image evaluation, indicating its effectiveness as a human-aligned evaluator for text-to-image generation.

2.2 Text-to-Image Generation Benchmarks

In terms of benchmarks, certain ones focus on assessing the images quality and alignment with human preferences (Kirstain et al., 2023; Li et al., 2024b; Xu et al., 2024; Yu et al., 2022), while others evaluate compositional generation by analyzing attributes like counting, color, and relationships (Huang et al., 2023; Park et al., 2021; Saharia et al., 2022). Recently, the focus has shifted to comprehensive evaluations (Lee et al., 2024; Li et al., 2024a; Wu et al., 2024b). For example, HEIM (Lee et al., 2024) comprehensively evaluates models across 12 distinct dimensions of capability. Furthermore, some benchmarks begin to emphasize the reasoning capability of T2I models. Commonsense-T2I (Fu et al., 2024) and PhyBench (Meng et al., 2024) focus on evaluating the multimodal commonsense understanding of generative models. However, the factuality of T2I models has not been sufficiently evaluated in the literature (Lee et al., 2024; Huang et al., 2024).

3 T2I-FactualBench Construction

In this section, we detail the construction process of T2I-FactualBench, a benchmark designed to evaluate the factuality of T2I models when generating images that rely on rich world knowledge.

3.1 Knowledge Concept Collection

The first step in constructing T2I-FactualBench involves collecting a set of knowledge-intensive concepts aimed to challenge T2I models by requiring generating precise visual details, rather than merely depicting general concepts. In this paper, we define a **Knowledge Concept** as a concept that has limited hyponyms in the knowledge base BabelNet (Navigli and Ponzetto, 2012).

Concept Category. We source to CNER (Martinelli et al., 2024) as the corpus to construct the knowledge concept set. CNER is a task designed for recognizing nominal concepts and named entities within a unified category space. It utilizes the completeness and broad semantic coverage of lexicographer files for nominal concepts in WordNet, along with the widely adopted semantic categories for named entities in OntoNotes, resulting in the establishment of 29 distinct categories. Among the 29 distinct categories, we focus on eight categories that can be grounded in real-world entities, such as *animal*, *artifact*, and *food*, while excluding abstract concept categories, such as *language*, *law*, and *discipline*.

Knowledge Concept Filtering. Given the training dataset of CNER, we begin by filtering out concepts in the eight categories and use SpaCy for lemmatization to obtain their standard lexical

Category	Subcategory	Examples	Num
ANIMAL	Mammals, Bird, Insect	*Bombay cat, Bombay cat, Keeshond dog, Aberdeen Angus, Damaliscus lunatus, Podilymbus, Crocodylus*	376
LOCATION	Landmark, Natural landform	*Kinderdijk, Leaning Tower of Pisa, Mont Saint-Michel, Oriental Pearl Tower, Butte, Danxia landform*	357
PLANT	Flower, Fruit, Tree	*Carnation, Balsam fir, Prunus armeniaca, Shumard oak, Syzygium, Coral bush*	312
ARTIFACT	Vehicle, Sports equipment, Musical instrument, Clothing, Tool	*Tesla Model 3, Ski pole, Kazoo, Pillbox hat, Chanel 2.55 flap bag, Sweater vest, DNA sequencer*	267
PERSON	Person	*Taylor Swift, Usain Bolt, Audrey Hepburn, Diane Keaton, Barack Obama, Mark Zuckerberg*	132
FOOD	Food	*Yakitori, Shrimp tempura, Macarons, Lasagna, Mooncake, Mapo Tofu*	131
EVENT	Event	*COVID-19 pandemic, Ratha-Yatra, World Rally Championship, Tour de France*	40
CELESTIAL	Celestial	*Horsehead Nebula, Mars, Mercury, Uranus, Enceladus*	14

Table 1: Category, Subcategory, Examples, and Number of knowledge concepts in T2I-FactualBench

forms. We then utilize BabelNet to gather relevant information and determine whether it satisfies our definition as a knowledge concept. Each concept is queried in BabelNet to gather relevant synsets, synonyms, categories, hypernyms, hyponyms, and images. We aim to select knowledge concepts with fewer than four hyponyms, hypothesizing that such concepts are more likely to exhibit distinct visual attributes, making them well-suited for thoroughly evaluating the knowledge capabilities of T2I models. This strategy may raise the concern on selecting uncommon concepts, thereby significantly influencing model performance. However, during the collection process, we find that the curated set of knowledge concepts includes both widely recognized concepts, such as "Husky dog," "British Shorthair cat," "bucket hat," "Macaron," and "Taylor Swift," as well as concepts that may be less commonly encountered in everyday contexts.

Ultimately, we curate a dataset of 1,600 knowledge concepts as a pool across eight domains, including *animals, artifacts, food, persons, plants, celestial bodies, events, and locations*. We detail the categorical distribution of knowledge concepts in Table 1. For the detailed concepts collection, see Appendix A.

3.2 Text-to-Image Generation with Knowledge-Intensive Concept

Building upon the collection of knowledge-intensive concepts, we propose a three-tiered text-to-image generation task to comprehensively evaluate the factual accuracy of T2I models.

T1: Single Knowledge Concept Memorization. We define the first level T2I generation task as Single Knowledge Concept Memorization (SKCM), which aims to assess whether T2I models can accurately generate a single knowledge concept, such as its specific visual attributes. Specifically, we utilize all concepts from the knowledge concept pool to construct task prompts using the following template: "An image of {Knowledge Concept}". Note that, in this task, the T2I model is permitted to determine the state or action of the knowledge concept.

T2: Single Knowledge Concept Instantiation. Next, we introduce the second-level T2I generation task, referred to as Single Knowledge Concept Instantiation (SKCI). SKCI advances T2I generation evaluation by measuring the model's ability to accurately instantiate knowledge concepts under reasonable conditions, such as depicting diverse actions for animals or varying attributes for objects. SKCI is designed to test the model's understanding of the intrinsic properties and behaviors associated with knowledge concepts, which are often difficult to explicitly articulate in prompts.

Specifically, in SKCI, we define three types of instantiation: $T = \{Action, Attribute, Scene\}$. *Action* is designed for animal, artifact, person, plant and food categories, which instantiates knowledge concepts with different actions. *Attribute* is tailored for categories such as animal, artifact, person, plant and food, instantiating these concepts with diverse states. *Scene* is designed for the location category, instantiating knowledge concepts through various

environmental conditions, including weather and time of day. A knowledge concept $c \in C$ is randomly sampled, along with an instantiation type $t \in T$ selected according to the concept's category. Given c and t, a powerful LLM M, such as GPT-4o, is prompted to sample one reasonable instantiation phrase p. For example, a phrase may be "chasing a ball" for knowledge concept "basset hound dog". Lastly, the reasonable phrase p_i is combined with the concept c to prompt the LLM M to produce a SKCI prompt S, which is used as the textual input for the T2I model.

T3: Multiple Knowledge Concept Composition with Interaction. Finally, we define the third-level T2I generation task as Multiple Knowledge Concept Composition with Interaction (MKCC), which is designed to evaluate a model's ability to simultaneously compose multiple knowledge concepts within a single image. MKCC assesses not only the common challenges encountered in general concept composition (Huang et al., 2023), such as adherence to prompts and seamless integration, but also both the implicit and explicit semantic relationships between different knowledge concepts.

Specifically, given two randomly selected knowledge concepts c_1 and c_2 from the concept set C, we first prompt GPT-4o to determine whether a significant size disparity exists between c_1 and c_2. If such a size discrepancy is identified, the LLM then proceeds to generate the prompt S. For example, given the prompt "A soccer ball and a baseball", the T2I model must correctly generate one image where the soccer ball is significantly larger than the baseball.

Next, if the two concepts do not show significant size discrepancy, we prompt GPT-4o to generate actions or attributes that can be used to instantiate each knowledge concept, following a process similar to that used in SKCI. This instantiated knowledge concept composition critically evaluates the ability of the T2I model to simultaneously represent the distinctive visual features of different knowledge concepts under various instantiations. Note that, in this task, the T2I model is granted the flexibility to determine the interaction between the instantiated knowledge concepts.

Lastly, we incorporate specific semantic relationships between knowledge concepts to further assess the model's ability to composite multiple knowledge concepts that interact with one another. Specifically, given two foreground (animal, plant, food, person, artifact) concepts, c_1 and c_2, and an optional background (location) concept, we first use GPT-4o to instantiate these concepts and then determine the interaction feasibility between the two foreground concepts, as well the suitability of these interactions occurring within the background concept. If appropriate, we use GPT-4o to generate one plausible interaction phrase, which is combined with the optional background concept to construct the prompt S.

Ultimately, we have constructed 3,000 prompts across all three levels. For detailed information of three-tiered framwork, see Appendix A

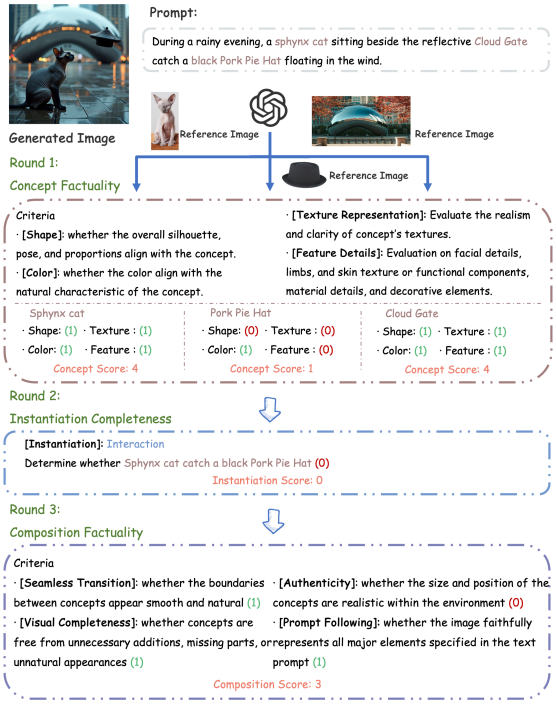

Figure 2: Multi-Round VQA based Factuality Evaluation Pipeline. We present an evaluation case in MKCC level.

4 Multi-Round VQA based Factuality Evaluation

To conduct an effective and efficient evaluation of the T2I model's performance on the proposed T2I-FactualBench, we introduce a multi-round visual question answering (VQA)-based evaluation framework, aided by advanced multi-modal LLMs. This framework consists of three VQA tasks: (1) Concept Factuality Evaluation; (2) Instantiation Completeness Evaluation; and (3) Composition Factuality Evaluation. Figure 2 provides an overview of the multi-round VQA-based evaluation for T2I-FactualBench.

4.1 Concept Factuality Evaluation

At the core of the T2I-FactualBench evaluation is the precise assessment of the factuality of the generated knowledge-intensive concepts. To achieve this, in the first round of VQA, we employ an advanced multi-modal LLM combined with the reference image as an effective proxy for the human evaluator to assess the factuality of the generated image. The reference image is obtained in the knowledge concept collection process sourced from BabelNet.

Specifically, given the knowledge concept c_i, its model generated image I_i, and the reference image I_i^R, we design a dedicated evaluation prompt to instruct GPT-4o to assess the factuality of the knowledge concept in the generated image across four dimensions: *shape, color, texture, and feature details*, as outlined in DreamBench++ (Peng et al., 2024). For the detailed definitions of the four dimensions, see Appendix B.1. For each dimension, GPT-4o assigns a score of 1, accompanied by a rationale, if the generated image meets the defined criteria. Otherwise, a score of 0 is assigned.

At the SKCM and SKCI levels, the generated image contains only a single concept. In contrast, at the MKCC level, the generated image encompasses multiple concepts. Therefore, we define the concept factuality score of I_i as:

$$\text{Concept Factuality} = \frac{1}{N_i}\sum_{j=1}^{N_i}\left(\frac{S_{ij}+C_{ij}+T_{ij}+F_{ij}}{4}\right) \tag{1}$$

where N_i denotes the number of concepts within I_i. $S_{ij}, C_{ij}, T_{ij}, F_{ij}$ represent the scores for the respective dimensions for the j-th concept in I_i.

4.2 Instantiation Completeness Evaluation

In addition to evaluating the factuality of the concept, we next evaluate whether the T2I model can produce precise instantiation of knowledge concepts.

We prompt GPT-4o to determine if c exists and the instantiation phrase p is successfully completed. If both conditions are met, the concept instantiation completeness score is assigned to 1. Otherwise, it is assigned a value of 0. For the detailed evaluation instruction, see Appendix B.2.

4.3 Composition Factuality Evaluation

Finally, in the last evaluation round, we design a VQA task to comprehensively evaluate the factuality of composing multiple knowledge concepts within a single image.

Specifically, given the text prompt t_i and the model-generated image I_i, we instruct GPT-4o to evaluate the composition factuality of the knowledge concepts in I_i across four dimensions: *Seamless Transition, Visual Completeness, Authenticity, and Prompt Following*. For the detailed definitions of the four dimensions, please see Appendix B.2. For each dimension, we apply the scoring system outlined in Section 4.1 and the composition factuality score of I_i is defined as:

$$\text{Composition Factuality} = \frac{S_i + V_i + A_i + P_i}{4} \tag{2}$$

where S_i, V_i, A_i, P_i representing scores for the respective dimensions for the i-th prompt, with each score taking a value of either 0 or 1.

5 Experiment

5.1 Experimental Setup

Text-to-image models We comprehensively evaluate the performance of seven text-to-image models on the T2I-FactualBench, including three variants of Stable Diffusion: (1) Stable Diffusion v1.5 (Rombach et al., 2022), (2) Stable Diffusion XL (Podell et al., 2024), and (3) Stable Diffusion 3.5 (Esser et al., 2024). Other models evaluated include (4) PixArt-alpha (Chen et al., 2024a), (5) Playground v2.5 (Li et al., 2024c), and (6) Flux.1 (Andreas Blattmann, 2024). For API-based models, we evaluate (7) DALL-E 3 (Betker et al., 2023)[1].

In addition to the aforementioned T2I models, we develop a visual knowledge injection method based on two subject-driven generation models:(8) SSR-Encoder (Zhang et al., 2023b), based on Stable Diffusion v1.5 and (9) MS-Diffusion (Wang et al., 2024b), based on Stable Diffusion XL. These models are capable of referencing one or more images during the generation process to enhance the factuality of subject representation. We also introduce a text-based knowledge injection method utilizing (10) Stable Diffusion 3.5* and (11) Flux.1 dev* due to the robust semantic comprehension capabilities afforded by their Diffusion Transformer (DiT) architecture. For further details about the methods, please refer to Appendix C.2.

[1]Note that due to policy restrictions on generating images of individuals, we did not evaluate DALL-E 3 for person knowledge concepts.

Model	SKCM	SKCI		MKCC		
	Concept	Concept	Instantiation	Concept	Instantiation	Composition
Text-to-image Generation						
SD v1.5	40.5	52.9	53.9	37.6	13.4	15.1
SD XL	45.8	59.9	65.8	51.7	28.0	35.4
Pixart	26.4	46.2	55.3	35.8	19.8	24.3
Playground	45.6	66.1	62.5	53.8	35.4	44.8
Flux.1 dev	35.6	54.9	58.0	56.9	54.1	63.8
SD 3.5	46.2	64.6	**71.2**	**68.9**	**59.2**	**75.5**
DALLE-3	55.5	72.4	88.5	71.3	70.2	85.6
Visual-Knowledge Injection						
SSR-Encoder	71.8 ↑ 31.3	69.0 ↑ 16.1	22.8 ↓ 31.1	43.1 ↑ 5.5	12.5 ↓ 0.9	9.5 ↓ 5.6
MS-Diffusion	**84.8** ↑ 39.0	**80.4** ↑ 30.5	50.8 ↓ 15.0	65.5 ↑ 13.8	32.9 ↑ 14.9	31.0 ↓ 4.4
Text-Knowledge Injection						
Flux.1 dev*	41.2 ↑ 5.6	60.3 ↑ 5.4	59.8 ↑ 1.8	64.2 ↑ 7.3	56.9 ↑ 2.8	72.6 ↑ 8.8
SD 3.5*	49.7 ↑ 3.5	66.7 ↑ 2.1	65.8 ↓ 5.4	67.9 ↓ 1.0	53.6 ↓ 5.6	64.7 ↓ 10.8

Table 2: **Main results** on T2I-FactualBench. We present the performance of text-to-image generation models and two distinct knowledge injection methods following Multi-Round VQA evaluation across three levels. We highlight the row of DALLE-3 in gray to denote the incompleteness of its evaluation data. **Model *** indicates that the model has undergone text-knowledge injection. ↑ and ↓ denote improvements and declines relative to their base models (SSR-Encoder→SD v1.5, MS-Diffusion → SD XL, Flux.1 dev* → Flux.1 dev, SD 3.5* → SD 3.5).

5.2 Evaluation Metrics

In addition to the proposed multi-round VQA-based factuality evaluation framework, we evaluate the T2I models using various metrics, including CLIP-T, CLIP-I (Radford et al., 2021), and DINO Score (Caron et al., 2021). For comparison, we also incorporate two MLLM-based evaluation methods, TIFA Score (Hu et al., 2023) and LLMScore (Lu et al., 2024b), which leverage MLLMs to provide a fine-grained assessment of text-image alignment.

5.3 Main Results

We first report the quantitative results of diverse text-to-image models on T2I-FactualBench.

Models Performance Across Three Levels. As shown in Table 2, the performance on T2I-FactualBench improves with the advancement of the backbone model. For example, Stable Diffusion 3.5 achieves higher concept factuality scores compared to previous models, such as SD v1.5 and SD XL. Furthermore, stronger models exhibit subtle changes in concept factuality scores when transitioning from SKCI to MKCC (e.g., SD 3.5: 64.6→68.9; Flux.1 dev: 54.9→56.9). They also perform better in composition evaluation (e.g., SD 3.5: 75.5; Flux.1 dev: 63.8). In contrast, weaker models experience significant declines in concept factuality scores (e.g., Playground: 66.1 → 53.8) and achieve lower composition factuality scores (e.g., Playground: 44.8).

Moreover, as instantiation complexity increases from SKCI to MKCC, all models exhibit a decline in Instantiation Completeness scores. These more intricate tasks require not only the retention of multiple knowledge concepts but also the ability to distinctly instantiate and effectively compose these concepts during generation. This observation highlights the limitations of existing T2I models in generating images involving knowledge concepts and their complex interactions.

We also observe a counter-intuitive trend that concept factuality scores tend to increase as task complexity increasing, from SKCM to MKCC. We hypothesize that this is due to the number of concepts varies across tasks of different levels. In fact, SKCM utilizes all the collected concepts. However, as described in Section 3.2, the prompts for SKCI and MKCC are constructed by randomly sampling one or more concepts along with an instantiation type, which are then provided to the LLM to generate a coherent and reasonable prompt. If a concept fails to generate a coherent and reasonable prompt, it will be discarded and resampled. Furthermore, at the MKCC level, the LLM is more likely to select combinations of more prevalent concepts. For example, the LLM may prefer combining concepts like *"Taylor Swift" and "Shiba Inu dog"* over *"Taylor Swift" and "flying lemur."* Therefore, SKCI and MKCC show higher concept factuality scores than

Method	Spear.	Kend.
CLIP-T	0.256	0.196
CLIP-I	0.235	0.175
DINO	0.277	0.207
TIFA Score	0.354	0.311
LLMscore	0.262	0.233
Concept Factuality (Ours)	**0.568**	**0.491**
- *w/o reference image*	0.442	0.376

Table 3: Correlation Scores between previous metrics and human evaluation in **Concept Factuality**. Spear. and Kend. represents Spearman and Kendall correlations, respectively.

SKCM. In Appendix D.2, we conduct an ablation study to validate this hypothesis.

Effect of Visual-Knowledge Injection. Table 2 also shows that the visual knowledge injection method (i.e., using reference images as visual knowledge) can significantly improve concept factuality of base models. However, their performance on instantiation and composition declines. We hypothesize that while visual knowledge injection enhances the factuality of concept generation, it simultaneously impairs the model's ability to follow instructions and accurately integrate multiple concepts.

Effect of Text-Knowledge Injection. In terms of the effect of text-knowledge injection, Flux.1 dev* shows significant improvements across various metrics with the additional textual descriptions of knowledge concepts. Conversely, SD 3.5* improves slightly in concept factuality but declines in instantiation completeness and composition factuality. This decline could potentially be attributed to the long prompts impairing instruction-following and concept composition capabilities. This finding suggests that while text-knowledge injection enhances concept generation accuracy, it may hinder instruction following with complex prompts.

5.4 Analyzes

Multi-Round VQA Metrics Better Aligning with Human Preference. We conduct experiments to investigate the effectiveness of the proposed multi-round VQA evaluation framework by comparing the VQA answers with human annotations. Specifically, we curate a validation set of 900 samples for all three level tasks and engaged three annotators on iTAG platform[2] to evaluate Concept Factuality, Instantiation Completeness, and Composition Factuality. The questions presented to the annotators are consistent with the prompts for Multi-Round VQA. We employ Spearman and Kendall correlations to quantify the alignment between human ratings and scores generated by Concept and Composition Factuality. For binary Instantiation Completeness scores, we compute accuracy. Detailed information about human annotations can be found in Appendix C.4.

Table 3 shows that our Concept Factuality evaluation aligns more closely with human judgments than previous metrics, highlighting its superior accuracy as reliable evaluation methods. Furthermore, when removing reference images from Concept Factuality assessment, there is a significant reduction in correlation coefficients, underscoring the necessity of providing reference images for accurate evaluation. In Appendix D, we conduct an inter-human agreement analysis and an error analysis for the Multi-Round VQA.

[2]https://www.alibabacloud.com/help/en/pai/user-guide/itag/

MLLMs	Concept		Composition		Instantiation
	Spear.	Kend.	Spear.	Kend.	ACC
Open-Source					
Qwen2.5-VL-7B	0.351	0.293	0.325	0.284	0.42
Qwen2.5-VL-72B	0.488	0.405	0.490	0.434	0.61
Internvl-2.5-78B	0.413	0.345	0.406	0.361	0.57
Closed-Source					
GPT-4o mini	0.347	0.283	0.444	0.392	0.63
Qwen-VL-Max	0.512	0.433	0.434	0.387	0.59
Gemini-1.5-Flash	0.541	0.470	0.513	0.466	0.73
Gemini-2.0-Flash	0.478	0.413	0.578	0.502	0.74
GPT-4o	**0.568**	**0.491**	**0.662**	**0.608**	**0.81**

Table 4: Comparisons of using different multi-modal LLMs as the backbone model for multi-round VQA evaluation. Spear. and Kend. represents Spearman and Kendall correlations, respectively.

T2I Models	Concept	Instantiation	Composition
SD v1.5	34.8 / 46.9	10.0 / 13.0	10.5 / 16.3
SD XL	52.9 / 70.3	13.0 / 16.0	28.8 / 38.5
Pixart	38.2 / 71.9	11.0 / 18.0	16.3 / 42.0
Playground	55.7 / 79.7	13.0 / 23.0	40.0 / 59.0
Flux.1 dev	55.3 / 84.3	30.0 / 46.0	54.8 / 82.3
SD 3.5	65.0 / 85.3	31.0 / 38.0	69.5 / 75.5

Table 5: Comparison of T2I models' performance on **Knowledge Concept** and **General Concept**.

Impact of Different MLLM for Multi-Round VQA Evaluation. To explore how different MLLM impact evaluation results, we test several closed-source models, including GPT-4o-mini, Gemini-1.5-Flash, and Qwen-Vl-Max (OpenAI, 2024; Team et al., 2023; Bai et al., 2023), alongside several open-source models such as Qwen2.5-VL-72B and Internvl-2.5-78B (Wang et al., 2024a; Chen et al., 2024b)

Figure 3: Concept Factuality Scores across 8 domains in the SKCM level for text-to-image models

As shown in Table 4, first, GPT-4o aligns more closely with human preferences in Concept and Composition Factuality and achieves the highest **81%** accuracy in Instantiation Completeness. Therefore, we choose GPT-4o as our evaluation model to ensure more accurate assessments. Second, in comparison, open-source models such as Qwen2.5-VL-72B (Bai et al., 2025) and InternVL-2.5-78B (Wang et al., 2025a) demonstrate competitive performance, highlighting the potential to explore trade-offs between performance, resource efficiency, and reproducibility in evaluation.

Ablating Knowledge Concept with General Concepts. To determine if models' poor performance on T2I-FactualBench is attributable to their limited factual generation capability related to knowledge concepts, we randomly select 100 prompts from the most challenging MKCC task and replace specific knowledge concepts with general ones (e.g., "Basset hound dog" to "dog"). Table 5 shows that models achieve significant improvements across three metrics when prompts including general concepts instead of knowledge concepts.

Models Face Challenge in Certain Domains. We analyze concept factuality scores across 8 domains within the SKCM level, as shown in Figure 3. Results indicate that models perform relatively well in animals, artifacts, and food, but poorly in plants and locations. We hypothesize that this discrepancy is due to the limited representation of plant and location concepts in the common training datasets (e.g., COCO and LAION). Further analysis of specific instances indicates that models struggle with generating concepts requiring detailed features. Specifically, when generating concepts related to plants and detailed elements of landmark buildings such as statues or architectural decorations, models fail to accurately capture intricate characteristics and complex textures.

6 Qualitative Analysis

Through an extensive qualitative analysis of images generated by various models on our benchmark, we identify four key deficiencies in current generative models: **Concept Error, Instantiation Failures, Realism Error, and Feature Mixture Error**. Detailed presentation of these findings is provided in Appendix D.6

First, models often obscure fine details of knowledge concepts or to generate similar but incorrect concepts. such as missing food concepts "Mapo Tofu" and "Gyoza" texture details, producing overly smooth surfaces and generating similar animal "wild boar" instead of "Baird's tapir". Second, models struggle with executing related attribute changes. For instance, they capture "ice skate" features but fail with variations like "broken ice skate". Third, when generating multiple objects, models often fail to integrate distinct features accurately, resulting in chaotic fusions that obscure individual characteristics. For instance, when generating both an "Egyptian Mau cat" and a "Basset Hound dog," the traits of these animals are often mixed inappropriately. Moreover, interactions among multiple concepts sometimes result in implausible scenarios, such as generating a "keeshond dog" embedded in a car window rather than positioned inside the car.

Additionally, with visual-knowledge injection, the output frequently lacks coherent integration of multiple knowledge concepts, appearing as simplistic concatenations. In text-knowledge injection, while improving factulity, it struggles with complex concepts like Mapo Tofu, failing to enhance instantiation performance or integration, indicating a need for refinement in handling complex tasks.

7 Conclusion

In this work, we introduce T2I-FactualBench, a benchmark to evaluate the factuality of knowledge-intensive concept generation as well as a multi-round VQA-based evaluation framework. Experiments on various T2I models show that current models still struggle with achieving high factuality in generating specific concepts and composing multiple concepts in one image.

Limitations

In this section, we discuss the limitations of this work: (1) **Knowledge concepts**. We only involve English knowledge concepts, which limits the comprehensiveness of evaluating the factual accuracy of text-to-image models on knowledge-intensive concepts. (2) **Generation task**. We propose a three-tiered generation task and design seven variations based on the intrinsic properties of knowledge concepts, including *memorization, action, attribute, scene, size, differentiating and interaction*. However, there are many more tasks relevant to real-world scenarios that we plan to explore in future work. (3) **Evaluation Scope**. The primary focus of our benchmark is to evaluate the factuality of text-to-image models when generating images that involve knowledge-intensive concepts. Consequently, we intentionally exclude other evaluation criteria, such as image fidelity and aesthetic quality, which, while significant, are beyond the specific scope of our study. We will provide a more comprehensive assessment of model performance in the future. (4) **Knowledge injection**. We explore two approaches for injecting external knowledge. However, both approaches come with their own limitations. We believe T2I-FactualBench will inspire further research on how to effectively inject external knowledge into T2I models.

Ethics Statement

During the collection of knowledge concepts, we rigorously eliminate any form of geographical or racial bias. Specifically, we employ random sampling from data sources and conduct strict manual curation to ensure a balanced concepts across different regions and ethnic groups. In the prompt generation phase, we implement keyword constraints to guide the model, ensuring that the content is reasonable and devoid of harmful elements. Additionally, we conduct a comprehensive manual review to further ensure the appropriateness and fairness of the content.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 62441617), and Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies.

References

Dominik Lorenz Andreas Blattmann, Axel Sauer. 2024. Flux.1.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Jinhe Bi, Yifan Wang, Danqi Yan, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. 2025. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *arXiv preprint arXiv:2502.12119*.

Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. 2024. Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. *arXiv preprint arXiv:2412.12359*.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024a. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *CoRR*, abs/2403.03206.

Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. 2024. Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? *arXiv preprint arXiv:2406.07546*.

Sebastian Hartwig, Dominik Engel, Leon Sick, Hannah Kniesel, Tristan Payer, Timo Ropinski, et al. 2024. Evaluating text to image synthesis: Survey and taxonomy of image quality metrics. *arXiv preprint arXiv:2403.11821*.

Wanggui He, Siming Fu, Mushui Liu, Xierui Wang, Wenyi Xiao, Fangxun Shu, Yi Wang, Lei Zhang, Zhelun Yu, Haoyuan Li, et al. 2024. Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis. *arXiv preprint arXiv:2407.07614*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

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

Hsin-Ping Huang, Xinyi Wang, Yonatan Bitton, Hagai Taitelbaum, Gaurav Singh Tomar, Ming-Wei Chang, Xuhui Jia, Kelvin CK Chan, Hexiang Hu, Yu-Chuan Su, et al. 2024. Kitten: A knowledge-intensive evaluation of image generation on visual entities. *arXiv preprint arXiv:2410.11824*.

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.

Nasiba Komil Qizi Jumaeva. 2024. Using the hyponyms for improving the beginners'vocabulary range. *Academic research in educational sciences*, 5(CSPU Conference 1):669–673.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2023a. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*.

Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. 2023b. Imagenhub: Standardizing the evaluation of conditional image generation models. *arXiv preprint arXiv:2310.01596*.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. 2024. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36.

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

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. 2024b. Evaluating and improving compositional text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5290–5301.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. 2024c. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *CoRR*, abs/2402.17245.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Tianwei Lin, Jiang Liu, Wenqiao Zhang, Zhaocheng Li, Yang Dai, Haoyuan Li, Zhelun Yu, Wanggui He, Juncheng Li, Hao Jiang, et al. 2024. Teamlora: Boosting low-rank adaptation with expert collaboration and competition. *arXiv preprint arXiv:2408.09856*.

Jiang Liu, Bolin Li, Haoyuan Li, Tianwei Lin, Wenqiao Zhang, Tao Zhong, Zhelun Yu, Jinghao Wei, Hao Cheng, Wanggui He, et al. 2024. Boosting private domain understanding of efficient mllms: A tuning-free, adaptive, universal prompt optimization framework. *arXiv preprint arXiv:2412.19684*.

Mushui Liu, Yuhang Ma, Zhen Yang, Jun Dan, Yunlong Yu, Zeng Zhao, Zhipeng Hu, Bai Liu, and Changjie Fan. 2025. Llm4gen: Leveraging semantic representation of llms for text-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5523–5531.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Wang, Yejin Choi, and Bill Yuchen Lin. 2024a. Wildvision arena: Benchmarking multimodal llms in the wild (february 2024). *URL https://huggingface.co/spaces/WildVision/vision-arena*.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2024b. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36.

Giuliano Martinelli, Francesco Molfese, Simone Tedeschi, Alberte Fernández-Castro, and Roberto Navigli. 2024. Cner: Concept and named entity recognition. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8329–8344.

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

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

OpenAI. 2024. Chatgpt. https://openai.com/index/gpt-4o-system-card/.

Jonas Oppenlaender. 2022. The creativity of text-to-image generation. In *Proceedings of the 25th international academic mindtrek conference*, pages 192–202.

Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. 2021. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. 2024. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

D She, Mushui Liu, Jingxuan Pang, Jin Wang, Zhen Yang, Wanggui He, Guanghao Zhang, Yi Wang, Qihan Huang, Haobin Tang, et al. 2025. Customvideox: 3d reference attention driven dynamic adaptation for zero-shot customized video diffusion transformers. *arXiv preprint arXiv:2502.06527*.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. 2024a. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024b. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Shanu Vashishtha, Abhinav Prakash, Lalitesh Morishetti, Kaushiki Nag, Yokila Arora, Sushant Kumar, and Kannan Achan. 2024. Chaining text-to-image and large language model: A novel approach for generating personalized e-commerce banners. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5825–5835.

Veera Vimpari, Annakaisa Kultima, Perttu Hämäläinen, and Christian Guckelsberger. 2023. "an adaptor-die type of situation": Perception, adoption, and use of text-to-image-generation ai by game industry professionals. *Proceedings of the ACM on Human-Computer Interaction*, 7(CHI PLAY):131–164.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. 2025a. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

X. Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. 2024b. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *CoRR*, abs/2406.07209.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. 2024c. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Yi Wang, Mushui Liu, Wanggui He, Longxiang Zhang, Ziwei Huang, Guanghao Zhang, Fangxun Shu, Zhong Tao, Dong She, Zhelun Yu, et al. 2025b. Mint: Multi-modal chain of thought in unified generative models for enhanced image generation. *arXiv preprint arXiv:2503.01298*.

Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. 2024a. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22227–22238.

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

Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. 2024b. Conceptmix: A compositional image generation benchmark with controllable difficulty. *arXiv preprint arXiv:2408.14339*.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Hao Jiang, Fei Wu, and Linchao Zhu. 2024a. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. *arXiv preprint arXiv:2404.14233*.

Wenyi Xiao, Zechuan Wang, Leilei Gan, Shuai Zhao, Wanggui He, Luu Anh Tuan, Long Chen, Hao Jiang, Zhou Zhao, and Fei Wu. 2024b. A comprehensive survey of datasets, theories, variants, and applications in direct preference optimization. *arXiv preprint arXiv:2410.15595*.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2024. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5.

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

Yuxuan Zhang, Jiaming Liu, Yiren Song, Rui Wang, Hao Tang, Jinpeng Yu, Huaxia Li, Xu Tang, Yao Hu, Han Pan, and Zhongliang Jing. 2023b. Ssr-encoder: Encoding selective subject representation for subject-driven generation. *CoRR*, abs/2312.16272.

Ziyu Zhao, Leilei Gan, Guoyin Wang, Wangchunshu Zhou, Hongxia Yang, Kun Kuang, and Fei Wu. 2024. Loraretriever: Input-aware lora retrieval and composition for mixed tasks in the wild. *arXiv preprint arXiv:2402.09997*.

A Dataset Details

A.1 Detailed Knowledge Concepts Collection Process

We elaborate on the meticulous process of collection and filtration employed for knowledge concepts pool.

The filtering process is outlined as follows: First, each concept is used as a query to retrieve relevant information from BabelNet, including the synset collection, categories for each synonym, hypernyms, hyponyms, and associated images. Because a single concept can have multiple synonyms (e.g., "Taylor Swift" may refer to the American singer or her eponymous album), we eliminate synsets that do not belong to the given category through keyword matching. To differentiate knowledge concepts from general concepts, we focus on selecting concepts with fewer than four hyponyms(Jumaeva, 2024). We hypothesize that such concepts are more likely to exhibit distinct visual attributes, making them well-suited for thoroughly evaluating the knowledge capabilities of T2I models.

A.2 Detailed Three-tiered Framwork

We present detailed information on the three-tiered framework of T2I-FactualBench in this section. In Table 6, we present our proposed three-tiered structure and seven tasks, including the number of prompts for each task and the evaluation method. In Table 9, we compare T2I-FactualBench to the knowledge domains of existing text-to-image benchmarks. Our T2I-FactualBench is the largest benchmark to date in terms of the number of concepts and prompts specifically designed to evaluate the factuality of knowledge-intensive concept generation.

B Multi-Round VQA Details

B.1 Concept Factuality Evaluation

In Concept Factuality Evaluation, we assess the factual accuracy of model's generated knowledge concept across four critical dimensions: *shape, color, texture representation, and feature details*. We provide a specific definition for each dimension in Table 10. Given that each category prioritizes distinct feature details, we design specific evaluation criteria for each category, as illustrated in Figure 5. The detailed Concept Factuality Evaluation prompt is provided in Figure 6.

Level	Task	Number	Evaluation
SKCM	Memorization	1600	Concept
SKCI	Action	200	Concept, Instantiation
	Attribute	200	
	Scene	200	
MKCC	Size	225	Concept, Instantiation, Composition
	Differentiating	225	
	Interaction	350	

Table 6: The three-level tasks in T2I-FactualBench

B.2 Instantiation Completeness and Composition Factuality Evaluation

In both the Instantiation Completeness Evaluation and Composition Factuality Evaluation, we first conduct a confirmation of presence, predicated on the existence of knowledge concepts within the generated images. If confirmed, the evaluation proceeds to the subsequent assessment stage; otherwise, the score is assigned 0. For the Instantiation Completeness Evaluation, we have tailored different evaluation prompts for each task, exemplified by the Size task prompt depicted in Figure 7.

In the Composition Factuality Evaluation at the MKCC level, we assess the composition accuracy of model's generated knowledge concepts across four critical dimensions: *Seamless Transition, Visual Completeness, Authenticity, and Prompt Following*. We provide a specific definition for each dimension in Table 10. Furthermore, we have crafted distinct prompts based on the number of knowledge concepts. The interaction variation necessitates an additional assessment of the background's composition factuality due to the presence of background knowledge concepts. Detailed prompts are illustrated in Figure 8 and Figure 9.

C Implementation Details

C.1 Model Details

We comprehensively evaluate the performance of 7 text-to-image models on T2I-FactualBench, including three variants of Stable Diffusion: (1) Stable Diffustion v1.5 (stable-diffusion-v1-5) , (2) Stable Diffustion XL (stable-diffusion-xl-base-1.0), and the latest (3) Stable Diffusion 3.5 (stable-diffusion-3.5-large) which incorporates the Diffusion Transformer (DiT) to enhance its capability in processing complex textual inputs. The (4) PixArt-alpha(PixArt-XL-2-1024-MS), leveraging Diffusion-Transformer technology, stands out for its minimal parameter footprint and expedited training process. Additionally, We include (5) Playground v2.5 (playground-v2.5-1024px-aesthetic), an advanced model evolving from Stable Diffu-

sion XL, which is fine-tuned to generate visually superior images that better resonate with human aesthetic preferences.. Furthermore, we evaluate (6) Flux.1 (FLUX.1-dev) , a successor to Stable Diffusion, featuring a novel hybrid architecture that merges multimodal processing proficiency with the parallelized functionality of the Diffusion Transformer. For API-based models, we evaluate the performance of (7) DALL-E 3 (Betker et al., 2023). When generating images, we default to leveraging the GPT model to enrich the input text prompts with additional details.

We adopted the default hyperparameters specified for each model by their respective authors.

C.2 Two Knowledge Injection Methods

Visual knowledge injection models We use the knowledge concept images from Section 3.1 as reference visuals to assist the models in image generation. We select two subject-driven generation models: (8) SSR-Encoder (Zhang et al., 2023b), based on Stable Diffusion v1.5, and (9) MS-Diffusion (Wang et al., 2024b), built upon Stable Diffusion XL.

Text knowledge injection models We augment the prompts by appending the definitions of the knowledge concepts, acquired in Section 3.1. This textual augmentation aims to enable the models to generate more precise representations of the concepts. We select (10) Stable Diffusion 3.5* and (11) Flux.1 dev* model due to the robust semantic comprehension capabilities afforded by their Diffusion Transformer (DiT) architecture.

C.3 Evaluation Details

In our Multi-Round VQA evaluation, we utilized the GPT-4o-0513 model. The final score for each VQA task is computed by evaluating the accuracy of generated image against established criteria.

We also compute established metrics. Specifically, we derived the CLIP-T metric by calculating the cosine similarity between the textual features and visual features extracted from the generated images utilizing CLIP (ViT-L/14). Moreover, for the CLIP-I and DINO metrics, we computed the cosine similarity between the feature sets of reference images and those of generated images by CLIP (ViT-L/14) and DINO (Dinov2-small), respectively.

For TIFA score, we employed GPT-4o-0513 to generate question-answer pairs based on text prompts, utilizing BLIP (blip-vqa-capfilt-large) VQA model to answer the questions with the generated image. The TIFA score was calculated as the accuracy of the answers produced by the VQA system. For LLMscore, we used QwenVL-2.5-72B model to generate image descriptions and employed GPT-4o-0513 to evaluate the alignment between the image description and the text prompt as LLMscore.

Notably, when evaluating the composition of multiple knowledge concepts in the MKCC, we calculate the CLIP-I and DINO scores by determining the cosine similarity for each individual knowledge concept and then averaging these values to obtain a composite score.

C.4 Human Evaluation Details

We conducted human evaluations on iTAG platform. Specifically, we engaged three annotators to evaluate Concept Factuality, Instantiation Completeness, and Composition Factuality to ensure the robustness of our assessments. The questions presented to the annotators were consistent with the prompts for GPT-4o, as depicted in Figures 6, 7, 8, and 9, to minimize bias introduced by question formulation. Figure 10, show the interface for human evaluation on Concept Factuality.

To assess the alignment between all evaluation metrics and human experts, we curate a balanced dataset comprising 300 concept validation samples, 300 composition validation samples, and 300 Instantiation validation samples. Each validation sample is evaluated by three annotators and we calculate the mean score from the three annotators' evaluations to ensure the reliability of the manual annotations.

D Additional Results

D.1 Results of Previous Metric Scores

In Table 7, we detail the CLIP-T, CLIP-I, and DINO scores on the T2I-FactualBench across three levels. Some results in the table are similar to those in Section 5.3.

However, our analysis reveals several critical limitations inherent in the current metrics:

- **Inadequate Assessment of Complex Instructions.** As a bag-of-words model, CLIP-T fails to accurately assess a model's ability to follow complex instructions. Specially, we observe that model performance appears to improve as instruction complexity increases,

Model	SKCM			SKCI			MKCC		
	CLIP-T	CLIP-I	DINO	CLIP-T	CLIP-I	DINO	CLIP-T	CLIP-I	DINO
Text-to-image Generation									
SD v1.5	31.0	75.2	38.4	31.3	76.0	44.7	33.0	67.0	26.6
SD XL	31.3	74.6	42.3	31.9	75.7	47.5	34.8	64.2	27.2
Pixart	27.8	68.0	30.2	29.8	71.5	40.1	32.9	64.1	26.3
Playground	30.8	73.6	42.4	31.6	73.8	47.1	35.0	64.9	29.3
Flux.1 dev	29.4	73.8	38.2	30.6	75.1	45.7	34.1	65.5	27.0
SD 3.5	**31.4**	76.9	43.4	**32.1**	76.6	47.7	**35.6**	67.3	29.8
DALLE-3	30.3	73.6	47.9	30.9	70.9	45.7	34.4	64.2	28.3
Visual-Knowledge Injection									
SSR-Encoder	30.4 ↓ 0.6	**86.6** ↑ 11.4	63.9 ↑ 25.5	28.9 ↓ 2.4	**87.2** ↑ 11.2	**68.8** ↑ 24.1	30.4 ↓ 2.6	**73.0** ↑ 6.0	36.5 ↑ 10.5
MS-Diffusion	31.0 ↓ 0.3	83.4 ↑ 8.8	**65.3** ↑ 23.0	31.2 ↓ 0.7	81.6 ↑ 5.9	64.1 ↑ 16.6	33.6 ↓ 0.8	69.9 ↑ 5.7	**37.1** ↑ 9.9
Text-Knowledge Injection									
Flux.1 dev*	29.8 ↑ 0.4	77.9 ↑ 6.1	44.7 ↑ 6.5	30.7 ↑ 0.1	77.1 ↑ 2.0	49.4 ↑ 3.7	34.3 ↑ 0.2	66.5 ↑ 1.0	29.0 ↑ 2.0
SD 3.5*	31.3 ↓ 0.1	79.6 ↑ 2.7	48.0 ↑ 4.6	31.7 ↓ 0.4	78.6 ↑ 2.0	51.0 ↑ 3.3	34.9 ↓ 0.7	68.3 ↑ 1.0	31.3 ↑ 1.5

Table 7: **Additional results** on T2I-FactualBench. We present the previous metric evaluation of text-to-image generation models and two distinct knowledge injection methods across three levels. We highlight the row of DALLE-3 in gray to denote the incompleteness of its evaluation data. **Model *** indicates that the model has undergone text-knowledge injection. ↑ and ↓ denote improvements and declines relative to their base models.

transitioning from SKCM to MKCC, which misrepresents the model's true capability in handling intricate instructions.

- **Inability to Distinguish Different Models.** The previous metrics do not effectively distinguish between models. Notably, the performance discrepancies between strong backbone models and weak backbone models are minimal. For instance, the transition from SD v1.5 to SD 3.5 on the SKCM level reveals negligible changes in scores (CLIP-T: 31.0 → 31.4; CLIP-I: 75.2 → 76.9; DINO: 38.4 → 43.4), underscoring the metrics' inadequacy in capturing model advancements.

- **Lack of Fine-Grained Evaluation.** The current metrics are insufficient for fine-grained evaluation of model capabilities. They fail to assess effectively the models' ability to integrate multiple knowledge concepts, thereby providing an incomplete metric of the models' performance in complex compositional tasks.

While TIFA Score and LLMscore leverage MLLM assessments to provide a fine-grained reflection of the alignment between text and image, they are **not ideally suited for evaluating the factuality of T2I models in generating knowledge-intensive concepts and their compositions**. TIFA primarily focuses on the presence of objects, the correctness of general objects, and the accuracy of object attributes. Without a reference concept image as input, it cannot accurately assess the factuality of knowledge-intensive concept generation. For example, TIFA may recognize the presence of a *"dog"* in an image but cannot determine if it is specifically a *"basset hound dog."*

Similarly, LLMScore relies on VLMs to generate visual descriptions, which tend to use classified concepts instead of knowledge-intensive descriptions. For example, a VLM might describe a generated image as containing a *"dog"* rather than a *"basset hound dog."* This limitation affects the thorough assessment of factuality for generated knowledge-intensive concepts.

D.2 Ablation Study on Concept Factuality Across Levels

In Table 2, we noticed that concept factuality scores tend to increase as task complexity increasing from SKCM to MKCC. We believe this counter-intuitive trend is due to the number of concepts varies across tasks of different levels. While SKCM features a diverse range of **1600 concepts**, SKCI and MKCC use only **400 and 550 common concepts**, respectively, leading to higher factuality scores.

To validate this hypothesis, we collect the concepts that appear in MKCC, denoted as C_M. For SKCM, we calculate the Concept Factuality scores only for those concepts also found in C_M. As shown in the Table 8, the results reveal that, when considering the same set of concepts, the scores tend to be higher at the less complex SKCM compared to the MKCC.

Model	SKCM*	MKCC
SD v1.5	60.4	37.6
SD XL	65.7	51.7
Pixart	49.3	35.8
Playground	70.2	53.8
Flux.1 dev	68.9	56.9
SD 3.5	76.2	68.9
DALLE-3	82.3	71.3
SSR-Encoder	88.1	43.1
MS-Diffusion	92.9	65.5
Flux,1 dev*	74.3	64.2
SD 3.5*	80.8	67.9

Table 8: Ablation Results of diversity models performance in Concept Factuality. SKCM* indicates the subset where knowledge concepts are also present in MKCC.

D.3 Results of Model Performance in Different Domains and Dimensions

In Figure 4, we provide a comprehensive analysis of the concept factuality scores for 11 distinct models across eight knowledge concept domains at the SKCM level.

In Table 11 and Table 12, we present the performance of models across various dimensions of Concept Factuality and Composition Factuality on T2I-FactualBench.

Our comparative analysis reveals that even SOTA models exhibit stronger capabilities in representing overall Color, Shape, and Texture, while fine-grained **Feature Details remain challenging** for generative models. For composing multiple concepts, models achieve higher scores in Seamless Transition and Visual Completeness but **struggle with Authenticity and Prompt Following**. We suppose this is due to the inherent complexity involved in maintaining realistic spatial arrangements and precisely interpreting and executing detailed textual instructions.

D.4 Results of Error Analysis For Multi-Round VQA

We each collect 50 error cases for Concept Factuality, Task Completeness, and Composition Factuality evaluation, where GPT-4o assessment differed from human annotations. In Table 13, we present a comprehensive breakdown of these error cases, categorized by the specific dimensions of discrepancies observed in each evaluation metric. Our statistical analysis reveals: In Concept Factuality evaluations, discrepancies often arise in Texture Representation (32%) and Feature Details (57%), because of the need for advanced visual feature capture and analysis capabilities. For Task Completeness, errors frequently occur with complex Size (24%), Differentiating (28%), and Interaction (30%) instantiations in MKCC, as the model must accurately distinguish between multiple concepts and assess their correct instantiation and interaction. In Composition Factuality, errors frequently occur in the Authenticity (56%), requiring strong spatial recognition skills and common world knowledge.

D.5 Results of Inter-Human Annotators Agreement Rates

We observe substantial agreement among the annotators on the validation set. Specifically, for binary evaluations (Instantiation Completeness), consensus was achieved between at least two annotators in **87%** of the cases, with all three annotators agreeing in **74%** of the cases. For Likert-scale evaluations (Concept Factuality and Composition Factuality), we calculated a Krippendorff's Alpha (Krippendorff, 2018) of **0.72**, indicating a good level of agreement for a subjective task of this complexity. These metrics underscore the reliability of our human evaluations.

D.6 Qualitative results

We present additional qualitative cases in Figure 11. We identify four key deficiencies in current generative models: Concept Error, Instantiation Failures, Realism Error, and Feature Mixture Error.

E Cost of T2I-FactualBench and Multi-Round VQA Evaluation

We provide the necessary cost details as follows: We used the GPT-4o-513 API to filter knowledge concepts and generate different phrases for each task, and then created prompts based on the knowledge concepts and phrases. This process required 5,600 API calls, costing approximately $30. In the multi-round VQA evaluation, we used the GPT-4o-0513 model for all three levels. The evaluation of each model required 6,350 API calls, costing approximately $45.

Benchmarks	Tasks In Prompt Construction									Knowledge Concepts		Evaluation
	Basic	Action	Attribute	Scene	Size	Differentiating	Interaction	Back-Foreground	Multi-Concepts(3)	Domain	Num	
PartiPrompt	✓	✓	✓	✓	✗	✗	✗	✓	✓	6	212	Human and CLIP
HEIM	✓	✓	✓	✓	✗	✗	✗	✓	✓	6	342	Human and CLIP
KITTEN	✓	✗	✗	✓	✗	✗	✗	✓	✗	6	322	Human and Automatic Metrics
T2I-FactualBench	✓	✓	✓	✓	✓	✓	✓	✓	✓	**8**	**1600**	**Multi-Round VQA**

Table 9: **Comparing T2I-FactualBench to the knowledge domain of existing text-to-image benchmarks.** T2I-FactualBench covers more essential tasks in prompt construction, including single knowledge concepts understanding(marked blue) and multi-knowledge concepts composition(marked purple) Moreover, the knowledge concepts in our benchmark across 8 diverse domains(animals, artifacts, food, persons, plants,locations, celestial, events). With a total of 1,600 knowledge concepts, it stands as the most extensive benchmark in knowledge field.

Dimension	**Definition**
Shape	*Assess whether the overall silhouette, pose, and proportions align with the common shapes associated with the concept.*
Color	*Assess whether the concept's color scheme and lighting conditions align with the natural or expected hues, saturation, and brightness characteristic of the concept.*
Texture	*Evaluate the realism and clarity of concept's textures, ensuring authentic representation in key areas, free from blurriness, pixelation, or artificial effects, to uphold realistic integrity*
Feature Details	*Evaluate the accuracy, completeness, and logical placement of the concept's features. Focus on facial details, limbs, and skin texture to ensure they align with the natural or expected representation of the concept.*
Seamless Transition	*Assesse whether the boundaries between concepts appear smooth and natural.*
Visual Completeness	*Evaluate if concepts are visually consistent and free from unnecessary additions, missing elements, or unnatural appearances*
Authenticity	*Assess whether the size and position of the concepts are realistic within the environment. For example, a car should be much larger than a husky, and neither should be in nonsensical positions, like floating unsupported.*
Prompt Following	*Evaluate the extent to which the image faithfully represents all major elements specified in the text prompt.*

Table 10: Definition of various dimensions in Concept Factuality and Composition Factuality evaluation.

Model	**SKCM**				**SKCI**			
	Shape	**Color**	**Texture**	**Feature Details**	**Shape**	**Color**	**Texture**	**Feature Details**
Text-to-image Generation								
SD v1.5	39.5	38.0	34.5	21.6	54.9	48.9	49.4	30.5
SD XL	50.7	49.4	47.5	35.6	57.7	58.0	54.8	38.3
Pixart	32.9	36.5	35.0	28.3	43.2	46.5	44.2	25.3
Playground	50.4	53.3	51.6	36.4	63.0	62.8	65.3	45.7
Flux.1 Dev	61.5	63.2	59.0	42.7	59.0	62.8	59.3	38.0
SD 3.5	66.4	66.5	65.2	55.3	62.5	60.8	59.9	42.5
DALLE-3	78.1	75.8	75.5	56.1	78.9	69.2	82.1	58.3
Visual-Knowledge Injection								
SSR-Encoder	48.0	49.0	45.5	29.5	81.8	72.7	73.3	48.2
MS-Diffusion	64.5	63.3	59.6	47.0	81.9	82.8	74.2	69.2
Text-Knowledge Injection								
Flux.1 Dev*	69.2	70.2	68.8	48.1	64.5	68.3	65.3	42.7
SD 3.5*	72.1	73.7	72.3	53.2	70.7	74.2	73.0	48.8

Table 11: Performance of models across various dimensions of **Concept Factuality** on SKCM and SKCI.

Model	MKCC–Concept Factuality				MKCC–Composition Factuality			
	Shape	**Color**	**Texture**	**Feature Details**	**Seamless**	**Visual**	**Authenticity**	**Prompt Following**
Text-to-image Generation								
SD v1.5	39.5	38.0	34.5	21.6	16.1	18.2	12.8	13.2
SD XL	50.7	49.4	47.5	35.6	37.2	38.8	34.3	31.4
Pixart	32.9	36.5	35.0	28.3	25.8	27.5	22.0	21.7
Playground	50.4	53.3	51.6	36.4	46.1	48.1	43.1	41.9
Flux.1 Dev	61.5	63.2	59.0	42.7	67.2	69.3	61.5	60.1
SD 3.5	66.4	66.5	65.2	55.3	77.0	80.1	74.3	70.6
DALLE-3	78.1	75.8	75.5	56.1	85.6	87.2	88.8	82.5
Visual-Knowledge Injection								
SSR-Encoder	48.0	49.0	45.5	29.5	9.4	12.0	7.4	9.0
MS-Diffusion	64.5	63.3	59.6	47.0	31.0	35.0	27.9	30.7
Text-Knowledge Injection								
Flux.1 Dev*	69.2	70.2	68.8	48.1	75.3	77.1	70.6	67.4
SD 3.5*	72.1	73.7	72.3	53.2	66.4	68.6	61.0	62.2

Table 12: Performance of models across various dimensions of **Concept Factuality** and **Composition Factuality** on MKCC.

Concept Factuality		Instantiation Completeness		Composition Factuality	
Error Type	**Percentage**	**Error Type**	**Percentage**	**Error Type**	**Percentage**
Shape	12%	Action	6%	Seamless Transition	12%
Color	8%	Attribute	10%	Visual Completeness	10%
Texture Representation	**32%**	Scene	2%	**Authenticity**	**56%**
Feature Details	**57%**	**Size**	**24%**	Prompt Following	22%
-	-	**Differentiating**	**28%**	-	-
-	-	**Interaction**	**30%**	-	-

Table 13: **Error Summary**. Breakdown of Error Cases in Concept Factuality, Task Completeness, and Composition Factuality.

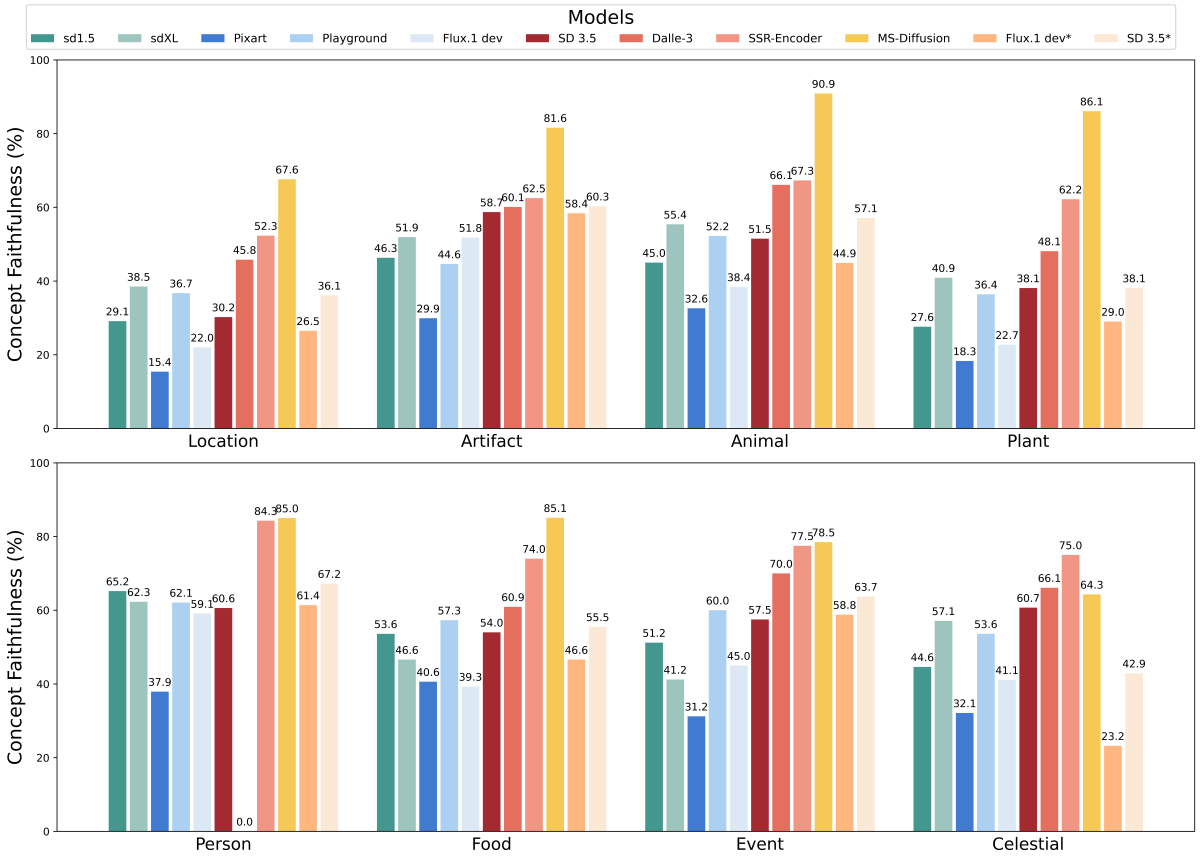

Figure 4: Concept Factuality Scores across 8 domains in the SKCM level for 11 Models.

Feature Details:

Animal: Evaluate the accuracy, completeness, and logical placement of the {concept}'s features. Focus on facial details, limbs, and skin texture to ensure they align with the natural or expected representation of the {concept}.

Person: Evaluate the accuracy, completeness, and logical placement of the {concept}'s features. Focus on facial details, limbs, and skin texture to ensure they align with the natural or expected representation of the {concept}.

Plant: Evaluate the accuracy, completeness, and logical placement of the {concept}'s features. Focus on facial details, limbs, and skin texture to ensure they align with the natural or expected representation of the {concept}.

Artifact: Evaluate the accuracy, completeness, and logical placement of the {concept}'s features. Focus on functional components, material details, and decorative elements to ensure they align with the natural or expected representation of the {concept}.

Food: Evaluate the accuracy, completeness, and logical placement of the {concept}'s features. Focus on functional components, material details, and decorative elements to ensure they align with the natural or expected representation of the {concept}.

Celestial: Examine the {concept}'s architectural details, unique decorations, and structural features, ensuring key elements and iconic details are precisely represented with correct proportions, symmetry, spatial layout and accurately reflect the original design.

Location: Examine the {concept}'s architectural details, unique decorations, and structural features, ensuring key elements and iconic details are precisely represented with correct proportions, symmetry, spatial layout and accurately reflect the original design.

Event: Evaluate the {concept}'s portrayal by examining the historical accuracy and representation of key figures, attire, and iconic scenes within the context of the event. Ensure that the visual narrative accurately depicts significant moments and the overall atmosphere, maintaining fidelity to documented historical accounts and cultural settings.

Figure 5: Feature details for eight knowledge concept categories.

Task Instructions:
You are provided with a concept image and a generated image. First, learn the concept of {concept} from the concept image. Then evaluate the {concept} in the generated image based on your own knowledge of {concept} and the concept of {concept} you learned from the concept image.
Concept Evaluation:
Evaluate whether the {concept} in the generated image matches the conceptual characteristics of the {concept} based on:
1. Shape Accuracy: Focus on the overall outline and structure of the generated {concept}. Assess whether the overall silhouette, pose, and proportions align with the common shapes associated with the concept.
2. Color Accuracy: Assess whether the generated {concept}'s color scheme and lighting conditions align with the natural or expected hues, saturation, and brightness characteristic of the concept.
3. Texture Representation: Evaluate the realism and clarity of the {concept}'s textures, ensuring authentic representation in key areas, free from blurriness, pixelation, or artificial effects, to uphold realistic integrity.
4. Feature Details: {feature_details}
Notice that: You should only consider the concept of {concept} in the generated image.
Evaluation Criteria:
Evaluate based on the criteria above. Score from 0 to 4 based on how many criteria are fully met:
- 0: None met
- 1: One met
- 2: Two met
- 3: Three met
- 4: All met
Explain your rating by providing a very brief core reason. Keep your explanation concise and to the point.
Input:
The generated image <image>. Caption of the generated image is {text}.
The concept image <image>. Caption of the concept image is {concept}.

Output Format(Notice that output must be without any bold formatting):
Total Rating: Your Rating
Shape Accuracy: 0/1 Reason:
Color Accuracy: 0/1 Reason:
Texture Representation: 0/1 Reason:
Feature Details: 0/1 Reason:

Figure 6: The prompt we used for Concept Factuality Evaluation with GPT-4o.

Task Instructions:
You need to analyze a generated image that contains two knowledge concepts: {type_1} {concept_1} and {type_2} {concept_2}. The task details are as follows:
1. Existence Check: First, confirm whether both concepts are present in the generated image. If either concept is missing, the Task Score should be 0.
2. Size Restoration: If both concepts are present, assess whether their size proportions in the image match their real-world counterparts.
3. Evaluation Criteria: Estimate the relative volume, height, or recognizable size ratio of each concept in the image to ensure they reflect the real-world size differences. If {concept_1} is indeed larger than {concept_2} as expected, the Task Score should be 1; otherwise, it should be 0.
Input Information:

The generated image: <image>.
Concepts: {type_1} {concept_1} , {type_2} {concept_2}
Output Format (Note: the output must be without any bold formatting):
Task Score:
Reason:

Figure 7: The prompt we used for Instantiation Completeness Evaluation with GPT-4o. A case of Size variation.

Task Instructions:
You are provided with a generated image, the text prompt, and two knowledge concepts: {type_1} {concept_1} and {type_2} {concept_2}. Your task is to evaluate the generated image based on these knowledge concepts and the text prompt.
1. Confirmation of Presence: determine whether both {concept_1} and {concept_2} are present in the generated image. If either is not present, the Total Rating should be 0. If both are present:
2. Integration Evaluation:
- Seamless Transition: Assess whether the boundaries between {concept_1} and {concept_2} are smooth, and ensure they integrate harmoniously with the surrounding environment.
- Visual Completeness: Evaluate whether {concept_1} and {concept_2} exhibit visual consistency without unnecessary additions, missing parts, or unnatural appearances.
- Authenticity: Assess if the size and position of {concept_1} and {concept_2} are realistic for the environment. For instance, a car should be much larger than a husky, and neither should appear in illogical positions, such as floating without support.
- Prompt Following: Evaluate whether the image faithfully represents all major elements specified in the text prompt and ensures that specific details, such as colors and shapes, are accurately depicted.
Evaluation Order: Please first confirm the presence of the concepts, then proceed to evaluate the integration.
Evaluation Criteria: Evaluate based on the criteria above. Total Rate from 0 to 4 based on how many criteria are fully met:
0: None met
1: One met
2: Two met
3: Three met
4: All met
Provide a concise explanation for the rating. Keep your feedback specific and to the point.
Input:
Generated image: <image>
Text prompt: "{text}"

Output Format (Notice that output must be without any bold formatting):
Total Rating: Your Rating
Seamless Transition: 0/1 Reason:
Visual Completeness: 0/1 Reason:
Authenticity: 0/1 Reason:
Prompt Following: 0/1 Reason:

Figure 8: The prompt we used for Composition Factuality Evaluation of **Two** knowledge concepts with GPT-4o.

Task Instructions:
You are provided with a generated image, the text prompt, and three knowledge concepts: {type_1} {concept_1}, {type_2} {concept_2} and background {type_3} {concept_3}. Your task is to evaluate the generated image based on these knowledge concepts and the text prompt.
1. Confirmation of Presence: determine whether {concept_1}, {concept_2} and {concept_3} are present in the generated image. If any of them is not present, the Total Rating should be 0. If all are present:
2. Integration Evaluation:
- Seamless Transition: Assess whether the boundaries between {concept_1}, {concept_2}, and background {concept_3} are smooth.
- Visual Completeness: Evaluate whether {concept_1}, {concept_2} and the background {concept_3} exhibit visual consistency without unnecessary additions, missing parts, or unnatural appearances.
- Authenticity: Assess if the size and position of {concept_1}, {concept_2}, {concept_3} are realistic for the environment. For instance, a car should be much larger than a husky, and neither should appear in illogical positions, such as floating without support.
- Prompt Following: Evaluate whether the image faithfully represents all major elements specified in the text prompt and ensures that specific details, such as colors and shapes, are accurately depicted.
Evaluation Order: Please first confirm the presence of the concepts, then proceed to evaluate the integration.
Evaluation Criteria: Evaluate based on the criteria above. Total Rate from 0 to 4 based on how many criteria are fully met:
0: None met
1: One met
2: Two met
3: Three met
4: All met
Provide a concise explanation for the rating. Keep your feedback specific and to the point.
Input:
Generated image: <image>
Text prompt: "{text}"

Output Format (Notice that output must be without any bold formatting):
Total Rating: Your Rating
Seamless Transition: 0/1 Reason:
Visual Completeness: 0/1 Reason:
Authenticity: 0/1 Reason:
Prompt Following: 0/1 Reason:

Figure 9: The prompt we used for Composition Factuality Evaluation of **Three** knowledge concepts with GPT-4o.

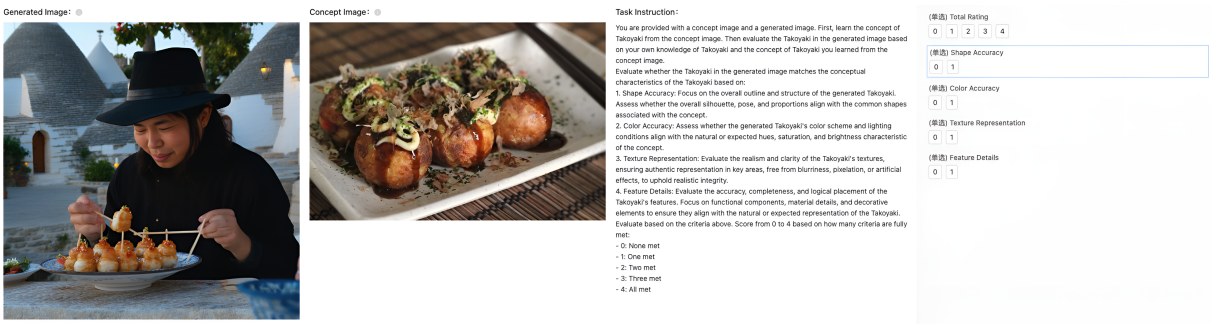

Figure 10: iTAG Interface for Concept Factuality Evaluation.

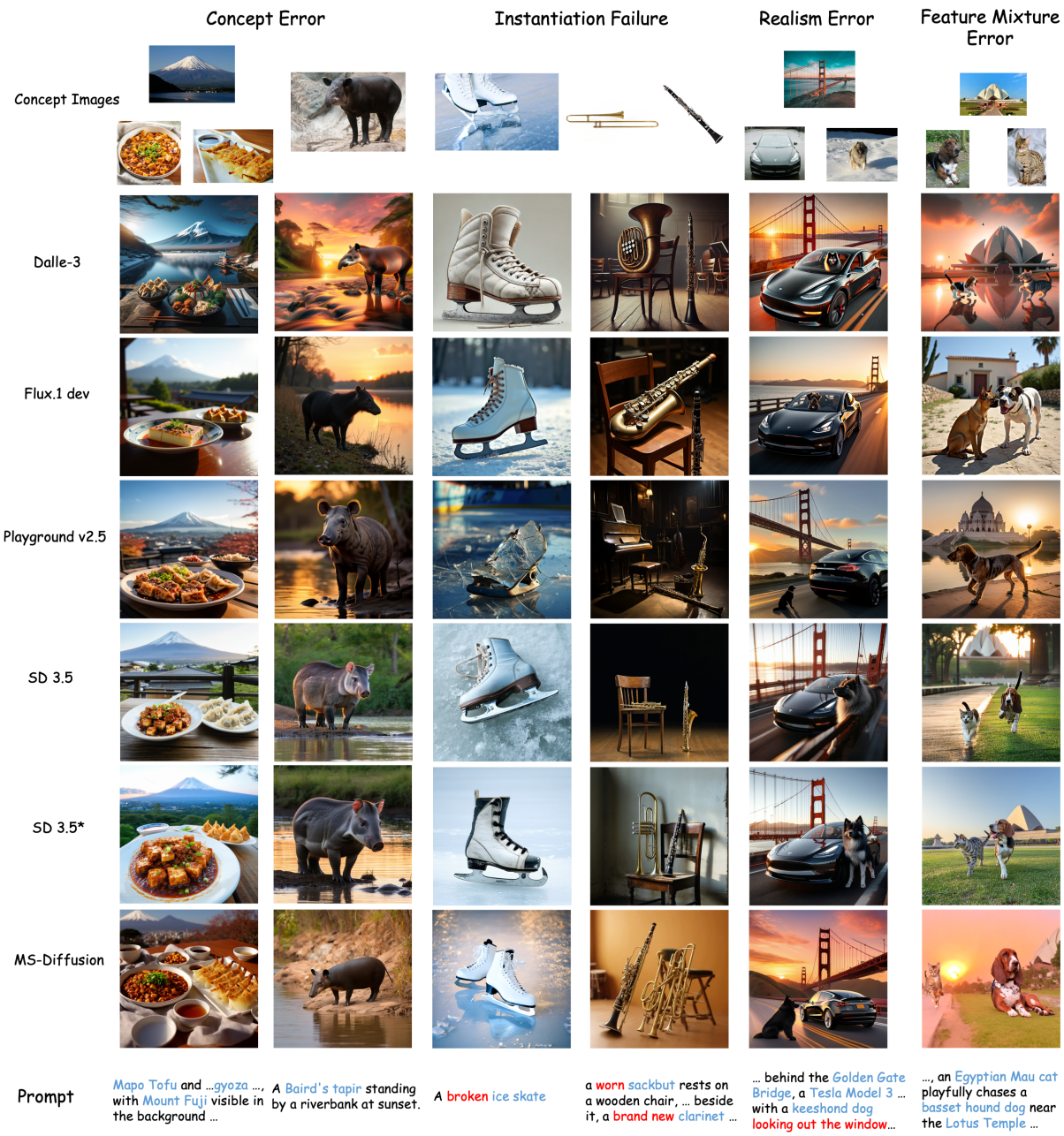

Figure 11: **Qualitative results**. Error cases of diversity models in T2I-FactualBench.