Hallucination as an Upper Bound: A New Perspective on Text-to-Image Evaluation

Seyed Amir Kasaei
Department of Computer Engineering,
Sharif University of Technology
a.kasaei@me.com

Mohammad Hossein Rohban
Department of Computer Engineering,
Sharif University of Technology
rohban@sharif.edu

Abstract

In language and vision–language models, hallucination is broadly understood as content generated from a model's prior knowledge or biases rather than from the given input. While this phenomenon has been studied in those domains, it has not been clearly framed for text-to-image (T2I) generative models. Existing evaluations mainly focus on alignment, checking whether prompt-specified elements appear, but overlook what the model generates beyond the prompt. We argue for defining hallucination in T2I as bias-driven deviations and propose a taxonomy with three categories: attribute, relation, and object hallucinations. This framing introduces an upper bound for evaluation and surfaces hidden biases, providing a foundation for richer assessment of T2I models.

1 Introduction

Hallucination, broadly defined as content generated from model priors rather than grounded in the input, is a critical challenge in contemporary AI. In large language models (LLMs) and vision–language models (VLMs), hallucination has been extensively studied because it undermines trust, factuality, and reliability, leading to a rich body of surveys and evaluation methods [8, 1, 3, 13, 14, 5, 11, 2, 9].

In contrast, hallucination in text-to-image (T2I) generative models has not been clearly framed. Current evaluations overwhelmingly treat the problem as one of *alignment*: verifying whether the objects, attributes, and relations explicitly mentioned in a prompt are faithfully represented in the generated image [15, 3, 6, 4, 7, 10, 12]. While these approaches have advanced prompt fidelity evaluation, they still capture only a *lower bound* of performance, focusing on what is requested, but overlooking a complementary dimension: *what does the model generates beyond the prompt?*

In this position paper, we propose a clearer framing of hallucination in T2I generation: rather than equating it with prompt misalignment, we define it as bias-driven deviations that manifest through unintended attributes, relations, or objects. This distinction is essential because alignment metrics establish only a *lower bound*, measuring whether requested elements are present, while hallucination evaluation sets an *upper bound*, capturing what the model introduces beyond the prompt. By explicitly recognizing hallucination as this complementary dimension, we expose hidden biases that remain invisible to existing evaluations and provide a principled foundation for more comprehensive and reliable assessment of T2I models.

2 Hallucination Taxonomy and Evaluation Perspective

We classify hallucinations in text-to-image generation into three types—object, attribute, and relation—based on how the model introduces unintended content. These hallucinations are distinct from

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.

alignment errors: instead of failing to follow instructions, the model extends the prompt beyond what the user explicitly requested. Each type raises unique challenges for evaluation and interpretability.

2.1 Object hallucination

Text-to-image models frequently introduce entities not mentioned in the prompt. We refer to this behavior as *object hallucination*. Unlike object neglect—where expected objects are omitted—hallucination involves generating new objects, driven by internal priors rather than user intent. Although these additions may be plausible, they often distort the focus or semantics of the prompt.

For example, a prompt like "a bowl of apples" might yield a bowl containing both apples and oranges; "a horse" might consistently appear with a rider; and "a street with cars" may include pedestrians or bicycles. These additions reflect model assumptions about scene completion, even when no such context was provided.

Formally, let a prompt P specify a set of objects $O = \{o_1, o_2, \ldots, o_n\}$. If the generated image contains a non-empty set O' such that $O' \cap O = \emptyset$, then O' constitutes object hallucination.

2.2 Attribute hallucination

Even when users omit attributes in their prompts, current text-to-image models often make specific visual assumptions. Rather than staying neutral, they tend to assign properties such as color, gender, style, or emotional tone by default. We define this behavior as *attribute hallucination*, distinguishing it from attribute misalignment, where the model misrenders a requested detail.

For instance, the prompt "a doctor" may consistently produce a male figure in a white coat, despite no gender or clothing instructions. A request for "a wedding cake" may always yield a tall, tiered, white cake, implicitly enforcing one cultural template. Similarly, "a child" might be shown smiling outdoors in polished clothing, reflecting idealized emotional defaults. These decisions, while plausible, reflect unprompted biases that reduce diversity and interpretive openness.

Let P refer to a prompt containing an object o with no explicit attributes. If the image includes an attribute a' not implied by P, we consider a' to be an instance of attribute hallucination.

2.3 Relation hallucination

Models are also prone to inserting relationships between objects—even when no such connection is described in the prompt. This behavior, which we term *relation hallucination*, introduces spatial or functional interactions that are not grounded in user input. It is distinct from relation misalignment, where a specified interaction is rendered incorrectly.

Such hallucinations may appear as default layouts or stereotyped activities. For example, "a man and a dog" may consistently depict the man walking the dog on a leash, implying control. A prompt like "a woman and a laptop" might always show her typing, suggesting a work scenario. Likewise, "a child and a book" often yields an image of the child reading, embedding an unintended learning narrative. These are not errors per se, but overlearned associations that compromise prompt neutrality.

Let P contain a set of objects $O = \{o_1, o_2\}$ with no explicit relation. If the resulting image includes a relation r not entailed by P, we categorize r as relation hallucination.

3 Conclusion

Our taxonomy reframes hallucination in text-to-image generation as a distinct, complementary dimension to prompt alignment. By identifying how models inject unintended objects, attributes, or relations, we highlight the need to move beyond alignment-based lower bounds and evaluate upper-bound behavior. Hallucination evaluation reveals model biases that undermine controllability, neutrality, and trust—factors critical for real-world deployment. We hope this perspective motivates new benchmarks that explicitly measure what models generate *beyond the prompt*.

References

[1] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, and M. Z. Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[2] K. Chen, Q. Chen, J. Zhou, Y. He, and L. He. Diahalu: A dialogue-level hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2403.00896*, 2024.

[3] Z. Chen, Y. Min, J. Zhang, B. Yan, J. Wang, X. Wang, and S. Shan. A survey of multimodal hallucination evaluation and detection. *arXiv preprint arXiv:2507.19024*, 2025.

[4] D. Ghosh, H. Hajishirzi, and L. Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[5] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.

[6] Y. Hu, B. Liu, J. Kasai, Y. Wang, M. Ostendorf, R. Krishna, and N. A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20030–20041, 2023.

[7] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu. T2i-compbench: A comprehensive benchmark for compositional text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[8] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[9] P. Kaul, Z. Li, H. Yang, Y. Dukler, A. Swaminathan, C. Taylor, and S. Soatto. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27228–27238, 2024.

[10] B. Li, Z. Lin, D. Pathak, et al. Vqascore: A vision-language qa approach for automatic text-to-image alignment evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

[11] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*, 2023.

[12] Y. Lim, H. Choi, and H. Shim. I-halla: Image hallucination evaluation with question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[13] H. Liu, W. Xue, Y. Chen, D. Chen, X. Zhao, K. Wang, L. Hou, R. Li, and W. Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

[14] P. Manakul, A. Liusie, and M. J. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

[15] Z. Qin, D. Cheng, H. Wang, H. Yi, Y. Shao, Z. Fan, K. Li, and Q. Lao. Evaluating hallucination in text-to-image diffusion models with scene-graph based question-answering agent. *arXiv preprint arXiv:2412.05722*, 2024.