Evaluating the Evaluators: Metrics for Compositional Text-to-Image Generation

Seyed Amir Kasaei[1] a.kasaei@me.com
Ali Aghayari[1] ali.aghayari@hotmail.com
Arash Marioriyad[1] arashmarioriyad@gmail.com
Niki Sepasian[1] sepasian.niki@gmail.com
MohammadAmin Fazli[1] fazli@sharif.edu
Mahdieh Soleymani Baghshah[1] soleymani@sharif.edu
Mohammad Hossein Rohban[1] rohban@sharif.edu

[1]Department of Computer Engineering,
Sharif University of Technology

Abstract

Text–image generation has advanced rapidly, but assessing whether outputs truly capture the objects, attributes, and relations described in prompts remains a central challenge. Evaluation in this space relies heavily on automated metrics, yet these are often adopted by convention or popularity rather than validated against human judgment. Because evaluation and reported progress in the field depend directly on these metrics, it is critical to understand how well they reflect human preferences. To address this, we present a broad study of widely used metrics for compositional text–image evaluation. Our analysis goes beyond simple correlation, examining their behavior across diverse compositional challenges and comparing how different metric families align with human judgments. The results show that no single metric performs consistently across tasks: performance varies with the type of compositional problem. Notably, VQA-based metrics, though popular, are not uniformly superior, while certain embedding-based metrics prove stronger in specific cases. Image-only metrics, as expected, contribute little to compositional evaluation, as they are designed for perceptual quality rather than alignment. These findings underscore the importance of careful and transparent metric selection, both for trustworthy evaluation and for their use as reward models in generation. Project page is available at this URL.

1 Introduction

Recent advances in multi-modal generative models, such as Stable Diffusion [26, 24, 6] and DALL-E [25], have made it possible to generate high-quality, natural, and diverse images from textual descriptions at low cost and on a large scale. Alongside this progress, a parallel line of research has emerged around a central question: *how can we faithfully assess the alignment between text and image, particularly in terms of compositional alignment?*

Text–image compositional alignment refers to how well the objects, attributes, and relations described in a prompt are reflected in the generated image. A variety of metrics have been introduced to evaluate this alignment. *Content-based metrics*, such as VQAScore [20], TIFA [11], DA Score [29],

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

DSG [5], and B-VQA [12], assess compositional properties through structured queries to the image. *Embedding-based metrics*, including CLIPScore [10], PickScore [16], HPS [31], and ImageReward [32], measure alignment via text–image similarity in a shared representation space or by using models trained on human preference data. *Image-only metrics*, such as CLIP-IQA [30] and Aesthetic Score [28], instead focus on perceptual quality and visual appeal, complementing alignment-based evaluation. Together, these metrics determine how models are evaluated and compared, and much of the field's reported progress is defined through their outcomes. Selecting appropriate metrics is therefore essential.

In addition to their role as evaluation tools, text–image alignment metrics have increasingly served as reward signals to enhance compositional generation in diffusion models through both inference-time and reinforcement learning strategies. For instance, ReNO [7] leverages multiple reward models, including CLIPScore [10], ImageReward [32], PickScore [16], and HPSv2 [31], to guide gradient-based optimization of initial noise during inference, improving both faithfulness and aesthetics. In contrast, ImageSelect [15] adopts a best-of-N sampling strategy, generating multiple candidate images from different initial noises and selecting the one with the highest *ImageReward* [32] score. On the training side, reinforcement learning methods like DPOK [8] fine-tune the diffusion model by maximizing ImageReward [32], using online policy gradient to strengthen compositional alignment.

Despite their widespread use in both evaluation and generation, there has been no thorough and comparative analysis of how well these metrics reflect human judgment. To address this gap, this work presents a comprehensive evaluation of 12 text–image compositional alignment metrics on 2400 generated text–image samples spanning 8 compositional categories [13], examining how closely each metric corresponds to human assessments.

2 Text-image Compositional Alignment

2.1 Evaluation Metrics

A range of metrics have been proposed for evaluating text–image alignment, each targeting different aspects of the correspondence. They can be grouped into three categories: (1) *embedding-based*, which rely on representations or preference models; (2) *content-based*, which use structured reasoning to assess compositional properties; and (3) *image-only*, which measure perceptual quality independently of text.

Embedding-based Metrics Embedding-based metrics evaluate alignment by comparing text–image representations in a shared multimodal space or by leveraging models trained on human preferences. A common baseline is *CLIPScore* [10], which measures cosine similarity between CLIP embeddings. Preference-supervised variants include *HPS* [31], which fine-tunes CLIP on human comparisons, and *PickScore* [16], which learns from pairwise preference judgments. *BLIP* [18] follows the embedding-similarity approach, comparing captions generated from images with the input text. Extending this idea, *ImageReward* [32] adds a reward head trained on ranked human preference data, capturing both textual relevance and perceptual quality.

Content-based (VQA-based) Metrics VQA-based metrics assess compositional alignment by casting text–image consistency as a question answering task. Questions derived from the prompt are posed to a pretrained VQA model, with scores based on the correctness of its responses. *VQAScore* [20] generates yes/no questions from the text, while *TIFA* [11] uses structured templates to cover objects, attributes, and relations. Variants target specific aspects: *DA Score* [29] asks entity–attribute questions to test binding, *DSG* [5] converts the text into a scene graph to verify entities and relations, and *B-VQA* [12] decomposes the text into object–attribute pairs, querying each with BLIP-VQA and combining the probabilities.

Image-only Metrics Image-only metrics assess perceptual quality independently of the prompt, providing complementary signals of realism and aesthetics. *CLIP-IQA* [30] predicts image quality by regressing CLIP embeddings against human quality annotations, while the *Aesthetic Score* [28] estimates aesthetic value from large-scale human ratings.

2.2 Benchmarks

Text–image compositional alignment benchmarks, such as T2I-CompBench++ [13], group prompts into categories that reflect major alignment challenges. These include: *entity existence* (all entities appear without omission or hallucination), *attribute binding* (objects match specified properties like color or shape), *spatial relations* (2D and 3D object arrangements, e.g., "a cube on a sphere" or "a ball inside a box"), *non-spatial relations* (functional interactions, e.g., "a man holding a guitar"), and *numeracy* (accurate object counts, e.g., "five birds on a branch"). Together, these categories provide a structured and fine-grained basis for evaluating compositional alignment.

3 Experiments and Results

3.1 Experimental Setting

Our analysis is based on T2I-CompBench++ [13], which provides curated prompts across attributes (color, shape, texture), spatial relations (2D and 3D), non-spatial relations, complex prompts, and numeracy. Each prompt is paired with images from multiple text-to-image models and annotated with human evaluation scores. All resources (prompts, images, and scores) come from the benchmark; our contribution is to analyze how evaluation metrics align with these annotations using outputs from SD v1.4, SD v2, Structured Diffusion [9], Composable Diffusion [21], Attend-and-Excite [2], and GORS [12].

We evaluate five embedding-based metrics (PickScore [16], CLIPScore [10], HPS [31], ImageReward [32], BLIP-2 [18]), two image-only metrics (CLIP-IQA [30], Aesthetic Score [28]), and five VQA-based metrics (B-VQA [12], DA Score [29], TIFA [11], DSG [5], VQA Score [20]), covering embedding similarity, perceptual quality, and VQA-style reasoning.

3.2 Correlation Analysis of Evaluation Metrics

We assess the reliability of reward models by correlating their scores on T2I-CompBench++ generations with human evaluations 3.1. Spearman correlations, reported in Table 1,vserve as the main measure, while Pearson and Kendall results are provided in Appendix A for completeness.

Per-Category Breakdown of Correlation Results Table 1 highlights that the strongest correlations differ substantially across categories, indicating that *no single metric dominates overall*. In the attribute group, DA Score leads on color (TIFA second), while ImageReward ranks highest on shape and texture (DA Score second). For relational cases, VQA Score performs best on 2D spatial (HPS second), whereas DSG leads in 3D spatial (HPS and BLIP-2 second). Non-spatial relations are best captured by HPS, followed by ImageReward. In complex prompts, VQA Score shows the strongest alignment, with TIFA second, and in numeracy, TIFA ranks first with ImageReward next. *Across all categories, image-only metrics (CLIP-IQA, Aesthetic) remain consistently weak*, underscoring their limited value for compositional alignment.

Table 1: Spearman correlation of evaluation metrics with human scores across compositional categories on T2I-CompBench++. The highest value in each category is shown in bold, and the second-highest is underlined.

Metric	Color	Shape	Texture	2D Spatial	Non-Spatial	Complex	3D Spatial	Numeracy
CLIP [10]	0.282	0.291	0.535	0.369	0.439	0.276	0.315	0.223
PickScore [16]	0.263	0.270	0.516	0.299	0.432	0.167	0.139	0.337
HPS [31]	0.219	0.440	0.601	<u>0.410</u>	**0.535**	0.270	<u>0.416</u>	0.471
ImageReward [32]	0.580	**0.520**	**0.734**	0.394	<u>0.512</u>	0.424	0.401	<u>0.484</u>
BLIP2 [18]	0.250	0.287	0.546	0.369	0.353	0.235	<u>0.416</u>	0.366
Aesthetic [28]	0.056	0.195	0.078	0.136	0.061	0.051	0.123	0.036
CLIP-IQA [30]	0.092	0.078	-0.001	0.088	0.082	0.027	0.098	0.068
B-VQA [12]	0.610	0.388	0.690	0.255	0.371	0.372	0.330	0.444
DA Score [29]	**0.772**	<u>0.463</u>	<u>0.711</u>	0.318	0.453	0.488	0.297	0.462
TIFA [11]	<u>0.684</u>	0.336	0.423	0.311	0.351	<u>0.519</u>	0.195	**0.526**
DSG [5]	0.599	0.388	0.628	0.328	0.470	0.411	**0.427**	0.469
VQA Score [20]	0.678	0.405	0.701	**0.533**	0.495	**0.638**	0.339	0.473

Broader Insights on Metric Performance Several broader insights emerge from these results. First, *no single metric achieves strong and consistent correlation across all compositional categories*, indicating that reliance on a single signal is insufficient. Second, despite its widespread use [26, 23, 27, 1, 17, 14, 3, 24, 4, 19, 22], *CLIP never ranks among the top metrics*, underscoring its limitations as a standalone measure. Third, embedding-based metrics, particularly ImageReward and HPS, frequently appear among the strongest. Fourth, while VQA-based metrics are competitive, *they are not uniformly superior and are occasionally outperformed by embedding-based approaches*. Finally, image-only metrics such as CLIP-IQA and Aesthetic remain consistently weak, as expected since they do not assess text–image alignment.

3.3 Beyond Correlation: Regression Analysis of Metrics

To complement the correlation study, we ran a regression analysis using human scores as the target and all metric outputs as predictors. A separate linear model was fit for each compositional category, with coefficients shown in Table 2. These coefficients reflect each metric's joint contribution, revealing shifts in importance across categories. Embedding-based metrics such as HPS and PickScore show strong positive contributions, while ImageReward and VQA-based metrics (DA Score, VQA Score, TIFA) also play key roles. In contrast, CLIP-IQA and Aesthetic often receive negligible or negative coefficients. Overall, embedding-based and VQA-based metrics provide complementary signals, and while the exact top metric can differ from correlation results, the same families tend to dominate—highlighting that no individual metric proves universally strong across categories.

Table 2: Regression coefficients for predicting human scores on T2I-CompBench++ categories. Highest values are bold, second-highest underlined.

Metric	Color	Shape	Texture	2D Spatial	Non-Spatial	Complex	3D Spatial	Numeracy
CLIP [10]	-0.290	0.000	-0.148	0.000	**0.649**	0.000	-0.444	0.000
PickScore [16]	**2.467**	0.000	**1.564**	0.000	0.000	0.000	-2.801	0.000
HPS [31]	-0.097	**0.761**	<u>0.475</u>	**1.143**	<u>0.629</u>	0.000	**2.048**	**1.277**
ImageReward [32]	0.193	<u>0.334</u>	0.300	0.197	-0.095	<u>0.223</u>	0.008	0.053
BLIP2 [18]	-0.420	0.000	0.264	0.000	0.594	0.000	<u>0.510</u>	0.000
Aesthetic [28]	0.229	0.000	0.260	0.000	0.023	0.000	-0.194	0.000
CLIP-IQA [30]	-0.017	0.000	0.014	0.016	0.000	0.000	0.100	0.000
B-VQA [12]	-0.092	-0.063	0.059	-0.264	0.035	0.000	0.071	0.059
DA Score [29]	<u>0.754</u>	0.132	0.236	<u>0.428</u>	0.000	0.000	-0.223	0.051
TIFA [11]	0.211	0.028	-0.015	0.003	0.061	0.214	-0.002	<u>0.204</u>
DSG [5]	-0.043	0.068	0.101	0.000	0.023	0.000	0.161	-0.045
VQA Score [20]	0.112	0.082	0.064	0.319	0.068	**0.246**	-0.003	0.092

3.4 Distribution Patterns of Metric Scores

Different families of evaluation metrics exhibit distinct distributional behaviors (Figure 1). Embedding-based metrics (CLIPScore, PickScore, HPS, ImageReward, BLIP-2) tend to produce mid-range scores. CLIPScore, HPS, and BLIP-2 are concentrated around values of 0.25–0.5, while ImageReward spans a broader range (0.15–0.8). Image-only metrics behave differently: Aesthetic is narrowly concentrated around 0.5–0.6, whereas CLIP-IQA spans a much wider range of 0.1–0.9. In contrast, VQA-based metrics (B-VQA, DA Score, TIFA, DSG, VQA Score) are strongly right-skewed and often saturate near 1.0, reflecting their quasi-binary nature. These patterns reveal two major concerns for evaluation. First, certain embedding-based metrics such as CLIP have a restricted value range, with many samples clustered around mid-level scores, which makes it difficult to distinguish quality differences and undermines their usefulness as evaluation measures. Second, VQA-based metrics are biased toward high values, often saturating near the upper bound, which reduces their ability to separate stronger candidates.

4 Conclusion

We studied how well current evaluation metrics capture human judgment in compositional text–to–image generation. Our analysis showed that no single metric is consistently reliable across categories: embedding-based methods (e.g., ImageReward, HPS) and VQA-based metrics (e.g., DA Score, VQA Score) each contribute, but with varying strengths. Regression confirmed that

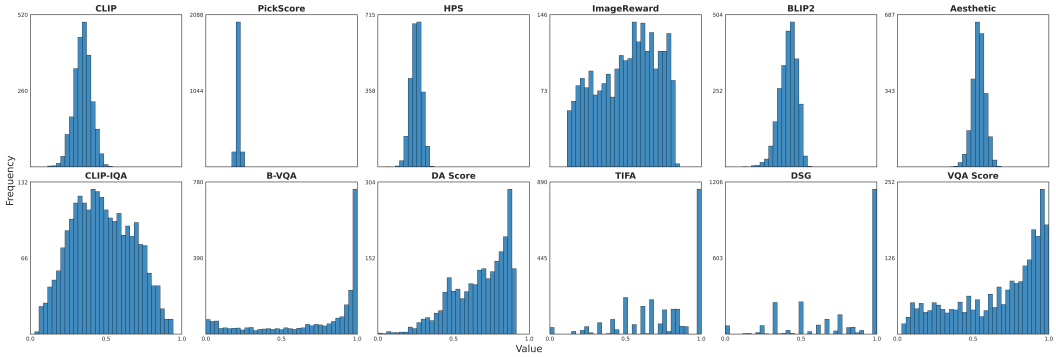

Figure 1: Value distributions of all analyzed metrics over T2I-CompBench++ generations (bin counts normalized by frequency).

their contributions shift when combined, while distributional patterns revealed limitations such as compressed mid-range scores in embeddings and saturation in VQA metrics. These results highlight the need for combining complementary metrics and for more faithful evaluation practices to guide progress in the field.

References

[1] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.

[2] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42:1 – 10, 2023. URL `https://api.semanticscholar.org/CorpusID:256416326`.

[3] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

[4] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, and Z. Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

[5] J. Cho, Y. Hu, R. Garg, P. Anderson, R. Krishna, J. Baldridge, M. Bansal, J. Pont-Tuset, and S. Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023.

[6] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[7] L. Eyring, S. Karthik, K. Roth, A. Dosovitskiy, and Z. Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Neural Information Processing Systems (NeurIPS)*, 2024.

[8] Y. Fan, O. Watkins, Y. Du, H. Liu, M. Ryu, C. Boutilier, P. Abbeel, M. Ghavamzadeh, K. Lee, and K. Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

[9] W. Feng, X. He, T.-J. Fu, V. Jampani, A. R. Akula, P. Narayana, S. Basu, X. E. Wang, and W. Y. Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=PUIqjT4rzq7`.

[10] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[11] 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*, pages 20406–20417, 2023.

[12] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 78723–78747, 2023.

[13] K. Huang, C. Duan, K. Sun, E. Xie, Z. Li, and X. Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[14] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10124–10134, 2023.

[15] S. Karthik, K. Roth, M. Mancini, and Z. Akata. If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection. *arXiv preprint arXiv:2305.13308*, 2023.

[16] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663, 2023.

[17] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023.

[18] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[19] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22511–22521, 2023.

[20] Z. Lin, D. Pathak, B. Li, J. Li, X. Xia, G. Neubig, P. Zhang, and D. Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024.

[21] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022.

[22] T. H. Nguyen and A. Tran. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7807–7816, 2024.

[23] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[24] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[25] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[27] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

[28] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[29] J. Singh and L. Zheng. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback. *Advances in Neural Information Processing Systems*, 36:70799–70811, 2023.

[30] J. Wang, K. C. Chan, and C. C. Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2555–2563, 2023.

[31] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023.

[32] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Appendix

A Correlation Study

In addition to Spearman correlation (Table 1 in the main text), we also report Pearson and Kendall correlation results in Appendix Tables 3 and 4. These complementary analyses confirm the same overall trend: no single metric consistently dominates across all categories, and the strongest performers vary depending on the type of compositional challenge. To further summarize these findings, Table 5 shows the frequency with which each metric appears among the top three in a category. This highlights the relative robustness of VQA Score and ImageReward, which each occur in six categories, followed by DA Score and HPS with four each. Importantly, the final set of consistently strong metrics includes two VQA-based methods (VQA Score, DA Score) and two embedding-based methods (ImageReward, HPS), suggesting that both types of evaluation are necessary for capturing the full spectrum of compositional alignment.

Table 3: Pearson correlation of evaluation metrics with human scores across compositional categories on T2I-CompBench++. The highest value in each category is shown in bold, and the second-highest is underlined.

Metric	Color	Shape	Texture	2D Spatial	Non-Spatial	Complex	3D Spatial	Numeracy
CLIP [10]	0.275	0.276	0.530	0.345	0.651	0.349	0.306	0.210
PickScore [16]	0.227	0.263	0.492	0.314	0.558	0.205	0.175	0.376
HPS [31]	0.240	0.457	0.555	<u>0.431</u>	**0.623**	0.355	**0.460**	<u>0.495</u>
ImageReward [32]	0.585	**0.555**	**0.737**	0.403	0.543	0.496	0.405	0.482
BLIP2 [18]	0.259	0.333	0.541	0.344	<u>0.620</u>	0.299	0.412	0.364
Aesthetic [28]	0.078	0.195	0.068	0.152	0.095	0.077	0.097	0.061
CLIP-IQA [30]	0.114	0.080	0.019	0.070	0.015	0.023	0.108	0.018
B-VQA [12]	0.621	0.348	0.644	0.237	0.462	0.426	0.335	0.460
DA Score [29]	**0.793**	<u>0.464</u>	<u>0.674</u>	0.324	0.519	0.513	0.298	0.485
TIFA [11]	<u>0.685</u>	0.364	0.415	0.298	0.451	<u>0.552</u>	0.212	**0.530**
DSG [5]	0.569	0.392	0.598	0.332	0.507	0.355	<u>0.441</u>	0.450
VQA Score [20]	0.682	0.409	0.617	**0.452**	0.542	**0.613**	0.348	0.487

Table 4: Kendall's τ correlation of evaluation metrics with human scores across compositional categories on T2I-CompBench++. The highest value in each category is shown in bold, and the second-highest is underlined.

Metric	Color	Shape	Texture	2D Spatial	Non-Spatial	Complex	3D Spatial	Numeracy
CLIP [10]	0.208	0.211	0.392	0.287	0.347	0.201	0.224	0.154
PickScore [16]	0.193	0.192	0.373	0.229	0.341	0.122	0.100	0.241
HPS [31]	0.157	0.326	0.441	<u>0.315</u>	**0.428**	0.201	<u>0.305</u>	0.346
ImageReward [32]	0.434	**0.388**	**0.549**	0.310	<u>0.408</u>	0.313	0.294	0.349
BLIP2 [18]	0.179	0.203	0.389	0.286	0.280	0.170	0.303	0.264
Aesthetic [28]	0.039	0.138	0.054	0.104	0.047	0.037	0.083	0.026
CLIP-IQA [30]	0.065	0.055	-0.002	0.068	0.063	0.018	0.068	0.045
B-VQA [12]	0.456	0.279	0.512	0.195	0.293	0.267	0.231	0.322
DA Score [29]	**0.603**	0.337	<u>0.534</u>	0.247	0.357	0.364	0.206	0.347
TIFA [11]	<u>0.559</u>	0.246	0.329	0.266	0.292	<u>0.405</u>	0.155	**0.400**
DSG [5]	0.499	<u>0.303</u>	0.503	0.292	<u>0.408</u>	0.325	**0.355**	<u>0.363</u>
VQA Score [20]	0.512	0.292	0.516	**0.422**	0.390	**0.481**	0.243	0.352

Table 5: Top-3 presence of each metric across various compositional categories. A ✓ indicates the metric is among the top 3 in that category. The last column shows the total number of categories where the metric appears in the top 3 based on spearman correlation.

Metric	**Color**	**Shape**	**Texture**	**2D Spatial**	**Non-Spatial**	**Complex**	**3D Spatial**	**Numeracy**	**Total**
CLIP [10]									0
PickScore [16]									0
HPS [31]		✓		✓	✓		✓		**4**
ImageReward [32]		✓	✓	✓	✓		✓	✓	**6**
BLIP2 [18]							✓		1
Aesthetic [28]									0
CLIP-IQA [30]									0
B-VQA [12]									0
DA Score [29]	✓	✓	✓			✓			**4**
TIFA [11]	✓					✓		✓	3
DSG [5]							✓		1
VQA Score [20]	✓		✓	✓	✓	✓		✓	**6**