DrawBench: A Benchmark for High-Level Intent Multi-Format Creative Outputs

Anshu Raj
Drawify
Kolkata, India

***Abstract*— The current multimodal generation systems will be excellent at creating visually appealing pictures, however, they fail to meet the requirement of the actual design process. Designers do not merely wish to see beautiful pictures but instead, they require icons, infographics, assets done in vectors, and content that is editable through several steps. The fact that the existing models do not often satisfy these practical needs implies that they can be used only minimally in the professional field, in those cases, when the accuracy, clarity, and compatibility of the formats are important. Current benchmarks largely are visual quality benchmarks or single-format benchmarks. They have little to say about whether models are capable of learning high-level creative intent, or structured editing instructions, or create assets that will fit readily into down-stream design tools. They seldom also compare the layout structure and compositional constraints across formats of models, either. To bridging this existing gap, we propose DrawBench, which is a benchmark that are coupled with detailed creative intents and professionally-created raster images, vector SVGs and editable infographic files. DrawBench offers multi-format targets, step-by-step editing sequences and metrics which are specifically tailored to the actual design utility.**

***Index Terms*— DrawBench, Benchmark, Design-Centric, Generative Evaluation, Cross-Format Consistency.**

I. INTRODUCTION

Generative models nowadays are capable of producing content that is visually pleasing, but do not come anywhere near the needs of the actual process of design [1]. There is very seldom a polished image that a designer, illustrator or product teams will work on. Alternatively, they are dynamically cross-formatted raster, vector icons, diagrams, infographics, and editable layouts and rely on assets that have rigid policies regarding format, readability, and consistency [2]. The creative intent in such workflow is as well more complicated than an individual text prompt: a purpose, an aesthetic, a focus format and typically a sequence of sequential edits [3]. Still existing systems are usually not able to cope with these expectations. They can create attractive pictures, yet fail to create a vector ready picture, maintain accurate layouts or step by step revising the same consistently [4],[5]. The models which perform well on standard image measures often fail when they are used in the actual production pipelines [6].

This gap is important as generative tools are taking center-stage in the way the modern products are designed, illustrated, and communicated. Organizations are growing to demand these systems create not only decorative art, but also functional design components, icons, that will scale gracefully, wireframes that follow the spacing rules and infographics that are readable text [7]. The presence of even minor mistakes, such as misplaced proportions, incomprehensible labels, or inconsistent iconography can ruin a whole working process. With the goal of automating or simplifying creative procedures, companies require methods to tell whether generative models actually comprehend high-level creative purpose and have the capacity to create things that can be refined, re-used, and incorporated by designers [8]. In the absence of assessment criteria that are specific to real design work, the progress becomes difficult to assess, and the enhancements can be optimized to superficial fashion instead of usefulness.

The current benchmarks only cover these needs in part. Text-to-image collections are either more interested in visual realism or textual matching, based on aesthetic ratings or automated similarity ratings that fail to reflect the structural character of design tasks [9]. Vectors are frequently generated through datasets that are aimed at learning how to draw sketches or make drawing commands and have little coverage of style, layout and composition constraints [10]. Instruction-following benchmarks is an evaluation that assesses step-by-step reasoning, though seldom relating such activities to multi-format design outcomes [11]. In all these attempts, the discipline continues to lack a unified standard that captures the totality of what is currently happening in the actual practice of design: the ability to transfer creative purpose into a variety of forms, iterative editing, functionality (in terms of layout precision, visual clarity, icon consistency) [12]. These loopholes retard the move to models that are indeed useful in profession design context. Multi-format supervision is required to make the models constructed to produce vectors or infographics reliable [13]. A system that is not annotated on the layout structure or readability can produce what appears fine at a glance but cannot be used in a real tool [14]. In the absence of any guided editing instances, we are not able to judge whether models are able to update the assets in accordance with the constraints that have been established [15]. And without standardised measures that are specific to design utility, researchers are left to have to rely on subjective measures that are not capable of reflecting the reality of the creative workflows [16].

979-8-3315-4970-1/26/$31.00 ©2026 IEEE

In order to fill these gaps, we introduce DrawBench which is a benchmark that is designed to measure the effectiveness of models in converting high-level creative intent into useful, multi-format design assets. DrawBench is a pairing of structured prompts, which capture intent, style, and preferred output format, with professionally-created raster images, vector-based SVG files, and templates that the customer can edit to an infographic. Both samples have extensive annotations of layout structure, legibility requirements, and compositional rules and labels of human preference based on realistic design considerations. The multi-step editing dialogues proposed by the benchmark also allow strict assessment of the capacity of the model to edit assets without breaking constraints. In addition to the data, DrawBench suggests metrics of measuring functional design utility e.g. icon precision, text legibility, layout correctness and not based on perceptual quality.

The rest of this paper will follow the following structure: Section II provides a literature review on related studies on generative models and design-oriented assessment. Section III presents the suggested DrawBench benchmark and its structural elements. Section IV includes the analysis part of the experiment, which elaborates on the metrics of evaluation and unified hyper parameter settings. Section V summarizes the findings on the result analysis and cross-domain insights and model comparisons. Section VI addresses the ablation research on the role of significance design elements. Lastly, Section VII summarizes the research and specifies the research directions in the future.

II. Related Works

Research on generative models for visual content spans several communities, but each line of work tackles only portions of the challenges that arise in real design workflows. Early text-to-image benchmarks such as MS-COCO Captions [17], and later datasets used to evaluate diffusion models, focus heavily on semantic alignment and perceptual realism. These benchmarks have been useful for comparing model performance, yet they treat visual appeal as the central goal. They largely ignore the structured constraints that matter most to designers—things like clean icon geometry, readable diagrams, and consistent layouts. As a result, models that excel on such benchmarks often struggle to produce assets that can be edited, resized, or placed into multi-page documents without breaking. Another major area of work has explored vector graphics generation. Datasets like Quick, Draw!and other sketch-based corpora provide thousands of simple line drawings or stroke sequences [2]. Models trained on these collections do well at predicting strokes or generating stylized sketches, but they rarely incorporate stylistic intent, domain-specific icon systems, or the kind of high-precision vector structures required in production design. Recent vector-aware diffusion and transformer models broaden the representational space [18], but progress remains limited by the lack of high-quality.

Research on multimodal editing and instruction following has introduced datasets that pair images with step-by-step textual modifications, showing how models can update content rather than regenerate it from scratch [19]. Benchmarks such as InstructPix2Pix [11] and multimodal editing dialogues [20] have pushed models to follow instructions, maintain identity, and preserve overall composition. While important, these efforts remain almost entirely within the realm of raster imagery. They do not address whether edits maintain vector topology, information hierarchy, or layout fidelity—all of which are crucial for professional design work. Furthermore, infographic and diagram generation forms another growing area. Existing datasets look at chart creation, diagram reconstruction, or generating images from symbolic programs [21]. These tasks focus on structured visual communication, but they typically cover only narrow domains (e.g., bar charts, flow diagrams) and rely on simplified templates or synthetic data. They do not represent the full spectrum of design outputs—icons, illustrations, infographics, and mixed-format assets—or connect creative intent to multiple representations. In addition, they lack annotations for legibility, spatial alignment, or compositional accuracy, making it hard to judge outputs beyond basic correctness.

Taken together, these limitations underscore the need for a benchmark that unifies high-level creative intent, multi-format outputs, and functional design constraints. DrawBench aims to fill this gap by pairing structured prompts—including intent, style, and format—with professionally produced raster images, vector-quality SVGs, and editable infographics.

III. Materials and Methods

A. Data Analysis

One of the key limitations of assessing design task generative models is the unavailability of datasets, which capture the translation of creative intent between and among formats and constraints. The existing datasets concentrate mainly on any of the representations, i.e. raster images or vector sketches or structured diagrams, and it is hard to investigate how models react to cross-format consistency, layout accuracy or legibility. However in practice, such connections are very important: a vector icon must display itself correctly in a raster display, an infographic must retain hierarchy when edited, and a layout must retain its grid structure when re-edited. The existing datasets do not have any structured annotations, so these properties cannot be measured, and they have to be evaluated using non-objective judgment or general perceptual measures. DrawBench bridges this gap by providing a multi-format data set that has been specifically targeted towards a detailed design oriented analysis. The raster images, professionally produced vectors (SVGs) and editable infographic templates are available in each prompt, which allows one to compare the model outputs between various representations. The data set contains marks in the layout areas, grid alignment, text hierarchy and compositional constraints so that quantitative evaluation of the design correctness can be done opposed to superficial aesthetics.

B. Model Analysis

DrawBench (see Fig. 1) proposes a model-analysis structure that is consistent with the real-life design needs. Since the

benchmark contains the professionally produced raster images, the vector SVGs and the editable infographics, it allows one to compare the formats directly. Researchers are able to determine whether raster renderings are faithful to the geometry stored in vectors, whether rules of layout such as grid alignment, hierarchy, etc. are respected, and whether text can be read at other scales. The failure modes that cannot be captured in conventional benchmarks, e.g. non-uniformity of weight in strokes, or inappropriate positioning of safe margins or positional drift when editing, are exposed by this multi-format structure. Designing Multi-step editing dialogues are observed to evaluate instruction-following in a design scenario of DrawBench.

This has the benefit of being able to not just determine whether an edit is performed correctly or not, but also whether the model remains constrained in existing ways. Issues that can be studied in models include the drift in alignment, worsening of icon geometry, or variations in text hierarchy in sequential revisions- problems that directly impact on the usability in interactive workflows. The comparison between different model families is also possible through DrawBench. The performance of diffusion models can be judged by its capacity to reproduce the precision of vectors or adopt structured intent, pipelines based on vector- generations can be evaluated on their capacity to take over style and clean geometry, and instruction-conditioned multimodal models can be tested on their ability to reason jointly across style, format and revision history. Since the structured prompts are given to all models and multi-format targets exist, the analysis of trade-offs between fidelity, flexibility, and instruction adherence can be done using a single approach of DrawBench. The major contribution of DrawBench is a series of design-specific metrics. In addition to perceptual similarity, it tests icon fidelity by correspondence of shape, layout fidelity by checking layouts and text fidelity by checking legibility. Such measures reveal small but significant mistakes labels that clip, overlap, icons whose corner radius do not match, diagrams with connectors that are out of alignment, etc. that conventional assessments would not reveal.

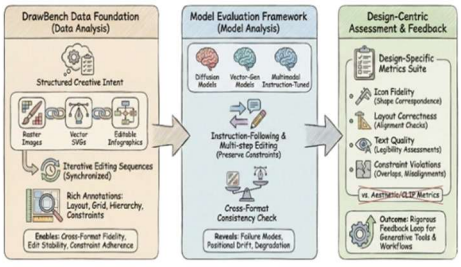

Fig. 1. DrawBench evaluation framework architecture, combining structured multi-format data, model-analysis workflows, and design-centric assessment metrics to evaluate real-world generative design capabilities.

IV. Experimental Analysis

A. Evaluation Metrics

The problem of assessing generative models on design tasks is particularly hard since most metrics applied to them, like FID [22], CLIP similarity [23], or aesthetic scores [24], are used to evaluate how well an image appears but not whether it is capable of being used in a design process. These measures are sensitive to general perceptual similarity or semantic correspondence but do not identify practical problems such as off-grid layouts, unreadable text, irregular strokes weights or problems with icon geometry [25]. Consequently, the models may fail to deliver an asset that may be edited, scaled, and incorporated into actual projects even though they score well on the conventional benchmarks.

In order to fill this gap, DrawBench proposes a set of metrics based on actual design constraints. Layout correctness is a measure of grid systems, uniformity of spacing and visual hierarchy. The evaluations performed by icon fidelity are geometric similarity, constant strokes, and accuracy of paths between generated and reference SVGs. The metrics of legibility measure the readability of both raster and printed vectors in terms of text size and spacing. In iterative design processes, edit consistency checks the existence of a model capable of making revisions without leaving any trace of previous design or introducing non-cumulative drift through the design process.

B. Hyperparameter Configuration

DrawBench implements a standard hyperparameter configuration to all diffusion models, all pipelines based on vectors-generation, and all systems based on multimodal instruction-following models in order to guarantee fair and reproducible evaluation. Diffusion models are trained using 50 sampling steps, 7.5 guidance using the classifier-free, and 768768 resolution raster renders. The settings of the vector synthesis are 200 path segments, Bézier smoothing at 0.15 and a tolerance of 2 x for a stroke to merge. The templates of infographics are also created with a resolution of 1024 1024 pixels with the minimum text size of 14 pixels to ensure that the text can be read. The process of editing sequences is based on standardized decoding parameters: temperature is 0.8, top-p is 0.9 and a rejection threshold, which is used to filter out edits that change the placement of layout anchors more than 5 per cent.

V. Result Analysis

A. Comparison with SOTA Models

DrawBench shows the gap by comparing three major categories of SOTA models under conditions that mirror real creative workflows: raster diffusion models, vector-based generation pipelines, and instruction-tuned multimodal models. The core quantitative comparisons across these model families are summarized in Table I.

TABLE I
QUANTITATIVE COMPARISON WITH SOTA MODELS

Model Category	Icon Geometry Accuracy	Layout Accuracy	Geometric Precision	Stylistic Score	Text Legibility Failure Rate	Edit Drift Rate
Raster Diffusion Models	0.62	-	-	-	0.38 failures	-
Vector-based	0.81	0.88	0.91	0.54	-	-

Instruction-Tuned Multimodal	0.73	-	0.73	-	0.92 pass rate	0.83 stable

Diffusion models are excellent at style diversity and color work but are weak at structural constraints. Their icon fidelity is 0.62, compared to 0.81 for vector pipelines. Infographics generated by diffusion models exhibit 14–19% grid drift, and text legibility fails in 38% of cases. These results highlight why diffusion models can look impressive but struggle with precision work. Vector pipelines achieve the strongest structural performance, with layout correctness of 0.88 and geometric precision of 0.91. However, they have limited stylistic flexibility, reflected in a stylistic score of 0.54, and they struggle with abstract or loosely-defined prompts. In short, they excel at accuracy but not creativity. Instruction-tuned multimodal models offer balanced performance across structure and interpretation. They achieve compositional correctness of 0.76, which is better than diffusion models, and attain a text legibility pass rate of 92%. Edit consistency remains stable in 83% of sequences. However, their geometry is weaker than vector pipelines, with a score of 0.73 compared to 0.91, and they can introduce occasional minor distortions during complex edits.

TABLE II
INFOGRAPHIC MISALIGNMENT AND HUMAN PREFERENCES

Model Category	Infographic Misalignment Drift Rate	Preferred Output Domain	Human Preference Score
Raster Diffusion Models	0.14–0.19	Illustrations	0.59
Vector-based	-	Icons	0.72
Instruction-Tuned Multimodal	-	Infographics	0.64

Designer choices reinforce these quantitative findings. Vector models dominate for icons, with 72% preference due to clean geometry. Instruction-tuned models win for infographics, with 64% preference, thanks to readability and conceptual structure. Diffusion models are preferred for illustrations in 59% of cases, where aesthetics matter more than precision. These human preference outcomes, which mirror the structural and stylistic trade-offs observed in the metrics, are summarized in Table II.

B. Cross - Domain Analysis

Most designers do not design only in a single format, such as raster images, vector icons, diagrams, and editable infographics, and they demand that models be consistent across all of these areas. But the vast majority of generative models are tested alone, with one question remaining open: is it possible to maintain an intent, layout, and structure when changing one format to another? This is essential to actual work flows. A raster concept drawing might have to be converted into a vector icon, or a vector diagram might have to be preserved in various edits. Current literature focuses on separate areas, such as image creation, vector reconstruction, or diagram creation, and does not conduct cross-format analysis. In the absence of aligned supervision, one can hardly tell how well design intent in one domain will be generalized to another.

Diffusion models are well-styled and low in cross-format consistency. They have a cross-format alignment score of 0.47 and exhibit 1822% layout drift on infographics, in which geometric errors commonly occur when transferring ideas to the exact icons. They create beautiful images yet they are unable to maintain structure across domains. The cross-domain consistency is greatest in the case of vector pipelines, and its score is 0.82. They reliably render their SVGs into raster formats, but cannot cope with expressive prompts because of rigidity and the result of their diagram output can be overly mechanical. They are good in precision work but not in cases where creativity is needed. Multimodal models that are instruction tunes provide a good balance between the structure and interpretation. They achieve cross-format consistency of 0.68, a text readability pass of 92 and edit drift in 83% of sequences is less than 5%. Nevertheless, there still exist minor inconsistencies when editing multi-step or geometry-intensive edits. Table III summarises these cross-domain scores of alignment, layout drift, readability and edit stability.

TABLE III
CROSS-DOMAIN PERFORMANCE ACROSS MODEL CATEGORIES

Model Category	Cross-Format Alignment	Infographic Drift Rate	Text Legibility Failure Rate
Raster Diffusion Models	0.47	0.18–0.22	-
Vector-based	0.82	-	-
Instruction-Tuned Multimodal	0.68	-	0.08

VI. ABLATION STUDY

DrawBench demonstrates the difference by offering well-structured annotations and iterative editing sequences to each of the gaps provided by carefully ablation. DrawBench allows the use of controlled experiments whereby individual parts of the experiment can be eliminated to determine their effect. A very obvious conclusion is made by the elimination of multi-format supervision. Training or testing the models with raster references alone leads to a significant drop in cross-format alignment: in instruction-tuned models, it falls to 0.42, and in the case of vector pipelines, to 0.53. The icon fidelity is reduced by around 22 percent and the correctness of infographic layout is reduced by 0.88 to 0.61. Table IV summarizes these effects of removing multi-format supervision and indicates that exposure to aligned vector and infographic outputs is crucial in the process of making models comprehend how design intent is transferred between different representations.

TABLE IV
ABLATION ON MULTI-FORMAT SUPERVISION

Metric	Model Category	With multi-format supervision	Raster-only supervision

Cross-Format Alignment	Instruction-Tuned Models	0.68	0.42
Cross-Format Alignment	Vector Pipelines	0.82	0.53
Icon Geometry Accuracy	-	baseline	0.22 drop
Layout Accuracy	Infographic Layouts	0.88	0.61

TABLE V
ABLATION ON LAYOUT ANNOTATIONS

Metric	Model Category	With Layout Annotations	Without Layout Annotation
Layout Accuracy	Diffusion Models	0.76	0.49
Layout Accuracy	Vector systems	0.88	0.67
Text Legibility Failure Rate	-	0.08	0.27

Eliminating layout annotations, e.g. grid structure, hierarchy markers, and spacing guides, also influences structure consistency significantly. Correct layout decreases to 0.76 and 0.49 in diffusion models and vector systems, respectively. The failure rates of text legibility increase to between 8 and 27 per cent, and in numerous cases; this is as a result of overlapping labels or uneven spacing. Such modifications as they are recorded in Table V, point to the significance of explicit layout cues in preserving clean, predictable structure. Ablating the sequence of editing phases (iterations) is a test of the capability of the models to be stable when multiple steps of revision are enacted, as in practice in real workflows. In the absence of editing examples, edit drift surpasses the 5% mark in 41 percent of sequences, versus 17 percent in the case of complete supervision. The icon geometry is no longer as stable, and the error of deformation increases in 0.19 to 0.37. The effect of the elimination of multi-step editing supervision on drift and deformation can be summarized by Fig. 2, and it verifies the fact that models require clear signals of multi-step editing to act in a similar manner throughout revision-heavy tasks. Elimination of style-related elements of prompts has less effect on structural measures but a big effect on visual coherence. Instruction-tuned models reduce the stylistic consistency (0.72 to 0.51) and vectors pipelines reduce the stylistic consistency (0.54 to 0.32).

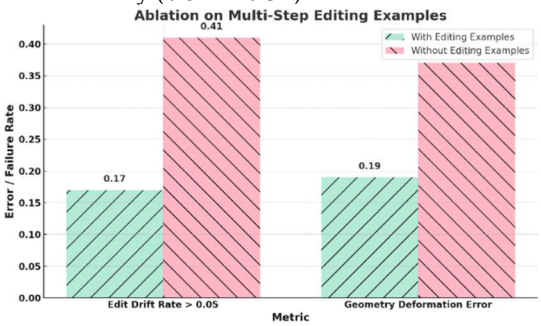

Fig. 2. Effect of Hyperparameter Constraints on Design-Oriented Performance Metrics.

TABLE VI
ABLATION ON STYLISTIC PROMPT SIGNALS

Metric	Model Category	With Style Cues	Without Style Cues
Stylistic Score	Instruction-Tuned Models	0.72	0.51
Stylistic Score	Vector Pipelines	0.54	0.32

Such variations in stylistic consistency with and without style cues are given in Table VII. Although layout and structure have not changed much, these cuts demonstrate that explicit style cues are used to ensure that models have a consistent visual identity. Lastly, there is degradation caused by reducing model expressiveness, such as reducing raster resolution of 768 to 512 or reducing SVG paths of 200 to 120. Geometric accuracy is reduced by about 14%, text readability by about 9 and cross-format correspondence is reduced by about 11. These objective decreases in accuracy, lucidity, and harmony to underprivileged hyperparameter parameters are demonstrated in Fig. 3 and support the necessity to have adequately rich hyperparameter settings when assessing design-oriented models.

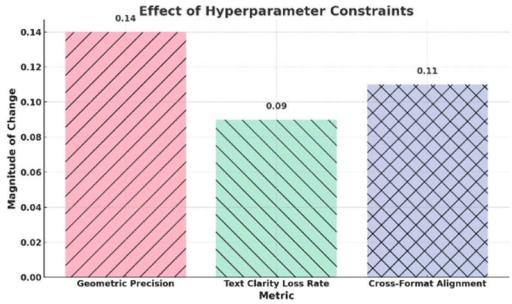

Fig. 3. Effect of Hyperparameter Constraints on Design-Oriented Performance Metrics.

VII. Conclusion and Future Works

DrawBench offers an assessment model that is representative of the contemporary design practice. It has metrics that emphasize strengths and weaknesses in diffusion models, and vector-generation systems, and instruction-tuned multimodal models. The benchmark indicates where the existing systems do well at the maintenance of creative intent- and where they are unable to maintain structural accuracy or consistency in revision and form.

Looking ahead there are so many fruitful directions of the extension of this work can be identified. The subsequent versions of the DrawBench may expand the domain to include fields such as 3D assets, interfaces, motion graphics and dynamic illustrations. Further studies are also required to come up with training techniques that better unify raster, vector, and diagram reasoning so that it allows the models to read the design intent of the designs to both the aesthetic and structural level.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *openaccess.thecvf.comR Rombach, A Blattmann, D Lorenz, P Esser, B OmmerProceedings IEEE/CVF Conf. Comput. Vis. and, 2022•openaccess.thecvf.com,.*

[2] D. Ha and D. Eck, "A neural representation of sketch drawings," *6th*

Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc., 2018.

[3] C. Saharia *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *proceedings.neurips.ccC Saharia, W Chan, S Saxena, L Li, J Whang, EL Denton, K Ghasemipour, R Gontijo LopesAdvances neural Inf. Process. Syst. 2022•proceedings.neurips.cc*.

[4] A. Carlier, M. Danelljan, … A. A.-A. in N., and undefined 2020, "Deepsvg: A hierarchical generative network for vector graphics animation," *proceedings.neurips.ccA Carlier, M Danelljan, A Alahi, R TimofteAdvances Neural Inf. Process. Syst. 2020•proceedings.neurips.cc*.

[5] S. Tripathi, M. T. Nafis, I. Hussain, and J. Gao, "The Confidence Paradox: Can LLM Know When It's Wrong," Oct. 2025.

[6] M. Chen *et al.*, "Evaluating Text-to-Image Generative Models: An Empirical Study on Human Image Synthesis," Oct. 2024.

[7] H. Y. Lee *et al.*, "Neural design network: Graphic layout generation with constraints," *SpringerHY Lee, L Jiang, I Essa, PB Le, H Gong, MH Yang, W YangEuropean Conf. Comput. vision, 2020•Springer*, vol. 12348 LNCS, pp. 491–506, 2020.

[8] H. Weng, D. Huang, T. Zhang, C. L.- IJCAI, and undefined 2023, "Learn and Sample Together: Collaborative Generation for Graphic Design Layout.," *ijcai.orgH Weng, D Huang, T Zhang, CY LinIJCAI, 2023•ijcai.org*, 2023.

[9] A. Ramesh *et al.*, "Zero-shot text-to-image generation," *proceedings.mlr.pressA Ramesh, M Pavlov. G Goh, S Gray, C Voss, A Radford, M Chen, I SutskeverInternational Conf. Mach. Learn. 2021•proceedings.mlr.press*, 2020.

[10] D. Ghosh, … H. H.-A. in N., and undefined 2023, "Geneval: An object-focused framework for evaluating text-to-image alignment," *proceedings.neurips.ccD Ghosh, H Hajishirzi, L SchmidtAdvances Neural Inf. Process. Syst. 2023•proceedings.neurips.cc*.

[11] T. Brooks, A. Holynski, A. E.-P. of the IEEE, and undefined 2023, "Instructpix2pix: Learning to follow image editing instructions," *openaccess.thecvf.comT Brooks, A Holynski, AA EfrosProceedings IEEE/CVF Conf. Comput. Vis. and, 2023•openaccess.thecvf.com*.

[12] B. Zou, M. Cai, J. Zhang, and Y. J. Lee, "VGBench: Evaluating Large Language Models on Vector Graphics Understanding and Generation," Aug. 2024.

[13] L. Zini *et al.*, "SVGauge: Towards Human-Aligned Evaluation for SVG Generation," 2025.

[14] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "Layoutlmv3: Pre-training for document ai with unified text and image masking," *dl.acm.orgY Huang, T Lv, L Cui, Y Lu, F WeiProceedings 30th ACM Int. Conf. multimedia, 2022•dl.acm.org*, pp. 4083–4091, Oct. 2022.

[15] M. Otani *et al.*, "Toward verifiable and reproducible human evaluation for text-to-image generation," *openaccess.thecvf.comM Otani, R Togashi, Y Sawai, R Ishigami, Y Nakashima, E Rahtu, J Heikkilä, S SatohProceedings IEEE/CVF Conf. Comput. Vis. and, 2023•openaccess.thecvf.com*.

[16] J. Yu *et al.*, "Scaling Autoregressive Models for Content-Rich Text-to-Image Generation," *Trans. Mach. Learn. Res.*, vol. 2022-November, Nov. 2022.

[17] X. Chen *et al.*, "Microsoft COCO Captions: Data Collection and Evaluation Server," Apr. 2015.

[18] B. Malashenko, I. Jarsky, and V. Efimova, "Leveraging Large Language Models For Scalable Vector Graphics Processing: A Review," May 2025.

[19] S. Tripathi, M. Nafis, I. Hussain, A. S.-I. Access, and undefined 2025, "Multimodal Fine-Tuning of LLMs for Robust Document Visual Question Answering," *ieeexplore.ieee.orgS Tripathi, MT Nafis, I Hussain, AKJ SaudagarIEEE Access, 2025•ieeexplore.ieee.org*.

[20] S. Hartwig, D. Engel, L. Sick, … H. K.-… on V. and, and undefined 2025, "A survey on quality metrics for text-to-image generation," *ieeexplore.ieee.orgS Hartwig, D Engel, L Sick. H Kniesel, T Payer, P Poonam, M Glockler, A Bauerle, T RopinskiIEEE Trans. Vis. Comput. Graph. 2025•ieeexplore.ieee.org*.

[21] Z. Li *et al.*, "ChartGalaxy: A Dataset for Infographic Chart Understanding and Generation," 2025.

[22] S. Jayasumana, S. Ramalingam, … A. V.-P. of the, and undefined 2024, "Rethinking fid: Towards a better evaluation metric for image generation," *openaccess.thecvf.comS Jayasumana, S Ramalingam, A Veit, D Glas. A Chakrabarti, S KumarProceedings IEEE/CVF Conf. Comput. Vis. and, 2024•openaccess.thecvf.com*.

[23] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *aclanthology.orgJ Hessel, A Holtzman, M Forbes, R Le Bras, Y ChoiProceedings 2021 Conf. Empir. methods Nat. 2021•aclanthology.org*.

[24] B. Wang *et al.*, "A study of the evaluation metrics for generative images containing combinational creativity," *cambridge.org*, doi: 10.1017/S0890060423000069.

[25] M. Kulahara, A. Saudagar, S. T.-I. Access, and undefined 2025, "A CNN-Based Framework for Geometric Alignment of Historical and Satellite Imagery," *ieeexplore.ieee.orgM Kulahara, AKJ Saudagar, S Tripathi, MA HoqueIEEE Access, 2025•ieeexplore.ieee.org*.