ChatGPT

Benchmarks for Text-to-Image Models (2022–2025)

In the past three years, a variety of benchmarks and evaluation frameworks have been proposed to assess text-to-image (T2I) generative models. Below, we review key efforts **year by year**, noting each benchmark's **year**, **main goal/target**, **evaluation type**, **notable evaluation axes**, and whether it uses **human judgments or automated metrics**. We then discuss the **current evaluation practices** and suggest **future directions** for evaluating T2I models.

2022: Early Benchmarks and Metric Suites

- **Multi-Task Benchmark (NeurIPS 2022)** – *Target:* Test fine-grained compositional capabilities of T2I models. This benchmark introduced **50 diverse tasks** (e.g. counting objects, binding attributes to correct objects, spatial relations, generating text in images) at easy/medium/hard difficulty [1]. *Eval type:* **Human evaluation** – experts rated images from different models (e.g. Stable Diffusion vs DALL·E 2) on each task [2] [3]. *Axes:* **Object count accuracy**, **attribute alignment**, **positional correctness**, **multi-object scene fidelity**, etc. *Human vs Auto:* Human (3,600 human ratings were collected) [2]. This revealed that while T2I models made progress, they **struggled with prompts containing many objects/attributes or requiring counting** [4].

- **TISE – "Bag of Metrics" (ECCV 2022)** – *Target:* Identify flaws in popular automatic metrics and propose a more reliable metric suite for T2I. This study noted problems with **Inception Score (IS)** (miscalibrated for single-object vs multi-object prompts) and **R-precision/Semantic Object Accuracy (SOA)** (overfitting to certain models), and that **FID** or CLIP-based scores alone often mis-rank models [5] [6]. *Eval type:* **Automatic metrics** – TISE introduced a combined set of metrics: an improved IS *(temperature-calibrated) and object-centric FID/IS for image quality, adjusted R-Precision and SOA for text-image alignment, plus new metrics for counting alignment and positional alignment in multi-object scenes* [7]. Axes: *Visual fidelity (IS*, FID), **semantic alignment** (CLIP/R-precision), **object correctness** (SOA), and **compositionality** (counting correctness, spatial arrangement) [6]. *Human vs Auto:* Automated (with validation against human rankings). Using this "bag of metrics" to rank models yielded **high consistency with human evaluation** and uncovered failures that single metrics missed [8]. The authors released the **TISE toolbox** to encourage fair, consistent evaluation [8].

- **DrawBench & PartiPrompts (2022)** – *Target:* Provide challenging prompt sets for human comparisons of T2I models. **DrawBench** (introduced with Google's *Imagen* model) is a curated list of ~200 prompts in 11 categories (e.g. complex spatial relations, rare objects, tricky compositions) used for **side-by-side human judgments** [9]. Similarly, **PartiPrompts** (Google's Parti model) offered prompts of varying complexity to test models' limits [10]. *Eval type:* **Human pairwise comparisons** of outputs from different models. *Axes:* Overall **image-text alignment** and **image quality** as perceived by humans across categories (e.g. counting, location of elements, novel concepts). *Human vs Auto:* Human. These early prompt-based benchmarks helped identify where models like DALLE-2, Imagen, etc. had qualitative strengths and weaknesses, but they provided **holistic judgments** only and lacked fine-grained error analysis [11]. (For instance, *DALL-Eval* (2022) used object detectors to check

if certain keywords appeared in the image, covering objects, counts, colors, spatial relations – but this binary approach missed many attributes like activities, material or style [12].)

2023: Fine-Grained and Human-Aligned Evaluations

- **Pick-a-Pic & PickScore (NeurIPS 2023)** – *Target:* Leverage **human preference data** to evaluate and improve T2I models. *Pick-a-Pic* introduced an **open dataset of 500K+ human comparisons** of images generated by different models for the same prompt [13] [14]. Real users of a web app were asked which of two images better matched their prompt (or if tied) [15]. From this data, the authors trained **PickScore**, a **learned scoring model** (reward model) fine-tuned from CLIP to predict human preference for a given image and prompt [16] [17]. *Eval type:* **Automatic learned metric** (PickScore model), validated by human studies. *Axes:* **Overall preference** (a composite of alignment, realism, aesthetics as implicitly judged by users). *Human vs Auto:* Automated (model outputs a score) but **grounded in human judgments**. PickScore achieved ~70.5% accuracy in predicting user choices (slightly **"superhuman"** vs individual human raters ~68%) [18] [19]. It correlated **much better with human rankings** than traditional metrics (e.g. Spearman 0.917 on MS-COCO prompts vs FID's – 0.900) [20]. The authors recommend using PickScore for more reliable evaluation of prompt adherence [20]. This work also suggested using **more user-relevant prompt sets** (instead of short captions) for evaluation [21] [22].

- **TIFA (ICCV 2023)** – *Target:* Provide an **accurate and interpretable "faithfulness" metric** using **question answering**. *Eval type:* **Automatic metric (VQA-based)**. TIFA (Text-to-Image Faithfulness Assessment) generates a set of **focused questions** from the prompt using an LLM, filters them for relevance, then uses a **Visual QA model** to answer those questions on the generated image [23] [24]. The fraction of correct answers yields a *faithfulness score* (how well the image content matches the prompt) [25] [26]. *Axes:* Fine-grained **object presence**, **attributes**, **relations**, etc., as captured by the questions (e.g. "Is there a red car? Is the dog on the left of the cat?"). It's **reference-free** and provides *which* aspects fail (via questions answered incorrectly) for interpretability. *Human vs Auto:* Automated (depends on pretrained QA/VQA models). TIFA showed a **much higher correlation with human judgments** on image-text alignment than CLIPScore or captioning-based metrics (Spearman $\rho \approx 0.60$ vs 0.33–0.34 for older metrics) [26] [27]. The authors released **TIFA v1.0 benchmark**: 4,000 diverse prompts (drawn from COCO, DrawBench, PartiPrompts, etc.), each with ~6 generated Q&A pairs (25K QA pairs total covering ~4.5K distinct prompt elements) [28]. This allows researchers to easily run standardized TIFA evaluations on their models [28]. *(Notable axes:* While TIFA's QA covers many properties, the authors note it inherently checks factual **presence/absence** of elements and could be extended to other aspects like counting or scene text in the future.)

- **ImageReward (NeurIPS 2023)** – *Target:* Introduce a **general-purpose reward model** for T2I alignment, trained on expert human comparisons. *Eval type:* **Automatic learned metric**, similar in spirit to PickScore but with a curated dataset. The authors collected ~137K high-quality human preference judgments focusing on alignment, image fidelity, and safety, and trained the **ImageReward** model on this data [29]. *Axes:* A weighted blend of **semantic alignment**, **image quality** and **"harmlessness"** (the human raters scored images for prompt alignment, realism, and absence of obvious flaws or offensive content [30]). *Human vs Auto:* Automated (reward model). In evaluations, **ImageReward outperformed other scoring models and metrics in matching human preferences**, making it a promising automatic metric for T2I [31]. This model is available openly [29]

and has been used to **fine-tune generation models** via reinforcement learning so that they produce more user-preferred outputs.

- **VPEval (NeurIPS 2023)** – *Target:* Provide an **interpretable, modular evaluation framework** via "Visual Programming." Rather than a single black-box metric, VPEval uses an LLM to orchestrate a *suite of vision expert modules* (object detectors, captioners, OCR, etc.) to evaluate different aspects of an image-step by step [32]. For example, given a complex prompt, it can call a face detector to verify the correct number of faces, a captioner to summarize the scene, a color detector to check an object's color, etc., and then combine these signals. *Eval type:* **Automatic multi-module pipeline**, with **textual+visual explanation** of scores. *Axes:* **Skill-specific checks** – e.g. **counting**, **spatial relations**, **object identity**, **attribute correctness** – each handled by an appropriate module [33] [34]. *Human vs Auto:* Automated. VPEval produces **human-readable evaluation reports** and was shown to correlate **more strongly with human judgments on targeted skills** than single-model metrics [32]. By decomposing evaluation, it avoids some failure modes of end-to-end metrics and increases transparency (users see which aspect failed). This approach highlights a trend of using **tool-augmented LLM "judges"** for T2I evaluation, improving reliability via cross-checking multiple criteria.

- **DSG (2023)** – *Target: DSG* stands for *"Davidsonian Scene Graph"*, an evaluation inspired by formal semantics. (This was introduced in late 2023 as an empirical framework.) It decomposes the prompt into a structured scene graph of entities and relations (akin to how a logician like Davidson would represent a scene) and then checks each part against the image [35]. *Eval type:* **Automatic QA-based**, similar to TIFA but using a fixed bank of fine-grained queries derived from the scene graph. *Axes:* **Fine-grained semantic categories** – objects, attributes, actions, cardinalities, etc. The method ensures a balanced coverage of different semantic aspects and provides *per-category scores*. *Human vs Auto:* Automated (with VQA). DSG offered high explainability but can be computationally intensive (many queries per image) [36]. In studies, DSG and related decomposition metrics contributed new information beyond CLIPScore and often correlated with each other (since they probe similar granular details) [37] [38]. Together with TIFA and VPEval, DSG reflects the 2023 push toward **breaking down "Did it get it right?" into multiple questions**.

- **Holistic Evaluation of T2I Models – HEIM (NeurIPS 2023)** – *Target:* Go beyond alignment and quality, and benchmark *many aspects* of generative models for a **comprehensive risk/benefit profile**. HEIM defined **12 evaluation aspects** critical for real-world use: **text-image alignment**, **image quality** (fidelity), **aesthetics**, **originality** (non-duplication and creativity), **reasoning** (can it combine concepts logically?), **knowledge** (factual or commonsense knowledge in outputs), **bias**, **toxicity**, **fairness**, **robustness** (e.g. to adversarial or out-of-distribution prompts), **multilinguality** (non-English prompts), and **efficiency** [39]. *Eval type:* **Mixed** – primarily **human evaluation** on curated test scenarios, with some automated measures for certain axes. The authors designed **62 scenarios** (prompt sets or tests), each targeting one or more aspects, and **evaluated 26 state-of-the-art models** across all aspects [40] [41]. For example, bias was evaluated by prompts involving different demographic descriptors; robustness by paraphrased or tricky prompts; efficiency by runtime and model size. *Axes:* The **12 aspects** listed above, making this the most wide-ranging benchmark to date [42]. *Human vs Auto:* Largely **human** – crowdsourced ratings or evaluator judgments were used for alignment, aesthetics, etc., and **expert review** for harms (with automatic filters as needed). The study found **no single model excels in all aspects** – different models have different strengths (e.g. one might be most **photorealistic** but another follows text instructions

better) [43]. By **open-sourcing the prompts, outputs, and human scores** [44], this benchmark enables the community to analyze trade-offs. HEIM underscores the importance of *holistic, multi-axis evaluation*: beyond making images look nice or match the caption, we must also evaluate safety, fairness, and other ethical dimensions of T2I generation.

2024: Improved Learned Metrics and Multi-Criteria Benchmarks

- **Human Preference Score v2 – HPSv2 (ArXiv 2024)** – *Target:* Build on human preferences at **greater scale and diversity**. HPSv2 introduced the **Human Preference Dataset v2 (HPD v2)** with **798K human comparisons on ~434K image pairs** – at the time, the largest such dataset [45] [46]. Unlike earlier datasets focused mostly on a couple of models, HPD v2 includes outputs from **9 different generative models** (covering diffusion, autoregressive, etc., plus real photos) to reduce bias [47]. Prompts were also cleaned (e.g. remove repetitive style tags) for fairness [48]. The team fine-tuned CLIP on this data to obtain **HPS v2**, an automatic scoring model predicting human preference [49] [50]. *Eval type:* **Automatic learned metric** (open-source model + a set of standardized prompts for evaluation). *Axes:* **Overall human preference** (like PickScore/ImageReward, it is a single score but implicitly combines alignment and quality as judged by people). *Human vs Auto:* Automated (learned from humans). HPS v2 showed improved generalization across datasets and model types, **outperforming prior metrics like HPS v1, ImageReward, and PickScore on correlation with human choices** [51]. Importantly, the authors studied **evaluation protocol** itself: using their data, they determined how many prompts are needed for stable model comparison, and released **four prompt lists** (≈200 prompts each) covering different styles ("Photo", "Concept Art", etc.) for a balanced evaluation [52] [53]. This gives a **standardized prompt suite** to test any new model with HPS v2, improving reproducibility.

- **EvalAlign (2024)** – *Target:* Fine-tune a multimodal model to **mimic detailed human annotation**. EvalAlign (by Tan et al., mid-2024) took a powerful vision-language model and **supervised-trained it on human-labeled data** to output **scores on multiple axes** for a generated image. Specifically, humans had given detailed labels for text-image pairs (e.g. scores for object fidelity, count accuracy, color correctness, style, action correctness, etc.). The tuned model can output an **interpretable score breakdown** mirroring these human criteria. *Eval type:* **Automatic (MLLM-based)** – essentially a learned evaluator that directly produces a **vector of scores** (one per criterion). *Axes:* **Faithfulness** to text overall, and fine-grained aspects like **objects present**, **counts**, **colors**, **styles**, **actions**, etc., as defined in the human instruction. *Human vs Auto:* Automated (model scoring), but closely aligned to human rating behavior (since it was trained on high-quality human judgments). This approach achieved **interpretable and stable evaluations** and was validated across 24 different T2I models. EvalAlign shows that with the right training data, we can **teach an AI to "think" like a human evaluator**, outputting multi-aspect scores in a single forward pass.

- **Positive-Negative VQA (PN-VQA) & EvalMuse Benchmark (late 2024)** – *Target:* Address bias in VQA-based metrics and create a large-scale benchmark with *element-wise* human annotations. *PN-VQA* (Han et al., 2024) modified the yes/no questioning approach: for each factual query about the image, it asks both the normal question *and* a negatively framed version (e.g. "Is there a cat?" / "Is there no cat?"). The metric combines these to counteract yes-saying bias and provide a more robust correctness score [54]. *EvalMuse-40K* is the accompanying **benchmark dataset of 40K prompts** with **human-annotated ground-truth answers for each prompt element** [55]. Humans marked, for each prompt, which parts the model got right or wrong, yielding both overall alignment scores and

per-element labels [55]. *Eval type:* PN-VQA is **automatic (VQA-based)**; EvalMuse is a **human-labeled test set** for evaluating metrics. *Axes:* **Fine-grained alignment** on **each described element** (object presence, count, attributes, etc.), measured in a robust way. *Human vs Auto:* PN-VQA is automated, but EvalMuse provides a **human-agreed "answer key"** to evaluate how well any metric or model is doing. This allowed researchers to rigorously compare metrics' accuracy – e.g. confirming that **no single metric excels universally** and that **combinations of methods work best for different types of prompts** [56] [57].

- **"What Makes a Good Metric?" (Ross et al., COLM 2024)** – *Target:* Critically **analyze recent automatic metrics** for text-image consistency. This study evaluated CLIPScore, TIFA, VPEval, and DSG against a checklist of desirable properties (e.g. *sensitivity* to changes in text or image, *robustness* to known shortcuts, *interpretability*) [58] [59]. It found that **none of the current metrics satisfy all criteria** and many have blind spots [60]. For example, some metrics were surprisingly insensitive to certain word changes, or relied on VQA models that tend to answer "yes" too often [61]. They also found TIFA, VPEval, and DSG, while more fine-grained than CLIPScore, **highly correlate with each other** and may not provide independent signals [37]. *Eval type:* Comparative study (using **human judgments as ground truth** to test metrics). *Axes:* Evaluated **metrics themselves** on axes like interpretability and robustness. *Human vs Auto:* Used human consistency ratings to see which metric best predicts them (none were perfect). The upshot is that 2024 saw a reflection on metrics: **metrics need to be scrutinized just as we scrutinize models**, to ensure we're measuring what we think we are.

2025: New Metrics, Domain-Specific Benchmarks, and Specialized Evaluations

- **cFréchet Distance (cFreD, 2025)** – *Target:* Integrate the text condition into image distribution comparison for **better fidelity+alignment metric**. cFreD (Koo et al., 2025) extends the popular FID metric by computing Frechet distance *conditioned on the text prompt*. Instead of comparing overall image features, it compares the distribution of features *for each prompt or conditioning*, thus penalizing cases where the image is high-quality but unrelated to the prompt [62]. *Eval type:* **Automatic metric**, mathematical extension of FID. *Axes:* **Joint quality & relevance** – it measures whether the model's output distribution, *given a prompt*, matches the real image distribution for that prompt (if such data exists) [62]. In experiments it achieved **very high correlation (~0.97)** with human rankings on the HPD v2 dataset [63]. *Human vs Auto:* Automated. cFreD is also model-agnostic and robust to new types of images (not fooled by unseen styles) [64]. This represents a push toward *distribution-level* evaluation that accounts for prompt semantics, unifying **image realism and prompt fidelity** in one metric [64].

- **CulturalFrames (CVPR 2025)** – *Target:* Evaluate whether models and metrics capture **cultural context and implicit expectations**. CulturalFrames introduced prompts that have literal descriptions plus culturally nuanced implications (e.g. symbols, attire, or activities expected in a given culture) [65]. Humans annotated whether outputs met both explicit and implicit expectations. Results were sobering: **metrics performed poorly** – e.g. they missed ~68% of explicit cultural mistakes and ~49% of implicit ones, and the best metric's correlation with human ratings was only ~0.32 [65]. *Eval type:* **Human benchmark** (difficult cases with cultural nuance). *Axes:* **Cultural alignment** – does the image match not just the words but the cultural context or norms implied.

Human vs Auto: Human (with analysis of metric failures). This benchmark highlighted the need for **culturally sensitive evaluation** and for metrics that go beyond surface-level matching, as T2I models start being used globally and can make subtle errors offensive or jarring to certain audiences [65].

- **GenomeBench (WACV 2025)** – *Target:* Establish a **structured human evaluation protocol** using **scene graphs** for fine-grained feedback [66] [67]. In this framework, each text prompt is parsed into a scene graph of **objects, attributes, relationships, and numerals** [68]. Human annotators are then guided through a *sequence of focused questions* about each node (e.g. *object node:* "How well is the *dog* rendered and does it exist as described?" on a 1–5 scale; *attribute node:* "Is the dog's fur color correctly depicted?"; *relation:* "Is the dog under the table as stated?") [69] [67]. Instead of binary yes/no, they give **graded ratings** that implicitly combine alignment and quality for that element [70]. These per-element scores roll up into an overall alignment/quality score with explanations [71]. *Eval type:* **Human (structured)** – uses human raters but in a systematic, rubric-based way. *Axes:* **Object correctness**, **attribute accuracy**, **relation accuracy**, and overall **image quality**, all on a per-element basis. *Human vs Auto:* Human. Advantages reported: much higher inter-annotator agreement and more consistent criteria, since complex prompts are broken into simpler sub-tasks [72] [73]. Also, the **fine-grained scores correlated better with holistic quality** than prior simple "aligned or not" labels [74]. GenomeBench provides **explainable feedback** – one can identify *which object* or *which attribute* dragged a model's score down, helping pinpoint model failure modes [71]. This is a promising direction for *human evaluations* to be more actionable and detailed.

- **Specialized Content Benchmark (MDPI 2025)** – *Target:* Rigorously evaluate T2I models on generating **images with embedded text and diagrams** (a known weakness of current models). Bosheah & Bilicki (2025) designed challenge prompts requiring **formatted or domain-specific content**: e.g. **mathematical equations**, **chemical structure diagrams**, **program code snippets**, **flowcharts**, **multi-line written text** (like a paragraph on a sign or a meme) [75]. These cases demand structural precision (correct symbols, legible letters, spatially arranged elements). *Eval type:* **Mixed** – They used **GPT-4** as an **automated evaluator** to score each output on multiple facets (text accuracy, readability, formatting, visual design, contextual relevance, etc.) [76] [77], and also gathered human evaluations for comparison [77]. *Axes:* **Text rendering accuracy** (are characters correct and readable), **structural layout fidelity** (e.g. alignment of equation elements, bond angles in molecule drawings), **semantic correctness** (does the content make sense given the prompt), and **error recovery** (if the model made partial mistakes, did it at least do so gracefully) [76]. *Human vs Auto:* Primarily automated (GPT-4 scoring), with human checks for alignment. The results showed **current models struggle heavily** with such tasks – e.g. incorrect math symbols, jumbled code syntax, incoherent multi-line text [78]. This benchmark establishes a **quantitative baseline for "text in image" generation**, an area likely to see focused improvements. It also demonstrates the use of an **LLM as an evaluator** for very complex criteria, which would be hard for simple metrics to capture.

- **Long-Prompt Benchmark – LPG-Bench (2025)** – *Target:* Test models on **long, detailed prompts (~250+ words)** and evaluate how well they capture *all* the described details. Wang et al. (2025) created **LPG-Bench** with 200 meticulously written prompts (averaging 250 words, often describing entire scenes or stories) [79] [80]. They generated 2,600 images from 13 leading models and collected **comprehensive human annotations** ranking how well each image matched its lengthy prompt [79] [81]. *Eval type:* **Human benchmark** (with a new automated metric proposed alongside). *Axes:* **Long-form alignment** – consistency with every part of a long description (which involves multiple objects,

interactions, background details, etc.). *Human vs Auto:* Human (rankings) for the benchmark; also introduced **TIT-Score** (Text→Image→Text Score) as an automatic metric: it uses a vision model to **describe the generated image in text, then compares that description to the original prompt** (via embedding or LLM) [82]. TIT-Score with an LLM comparison achieved a **7.3% absolute improvement** in pairwise accuracy on long prompts over prior metrics [82]. The overall finding is that long prompts pose a unique challenge – models often drop or alter details – and new metrics like TIT-Score can help quantify improvements. LPG-Bench fills an evaluation gap for **complex narrative or multi-sentence prompts**, which are increasingly relevant (e.g. users providing detailed scene instructions).

- **Image Editing Benchmarks (IE-Bench 2025)** – *Target:* Evaluate **text-guided image editing** specifically. Sun et al. (2025) proposed *IE-Bench*, focusing on cases where a model must **modify a given source image** according to a text instruction (e.g. "remove the glasses from the person" or "change the house's roof to red"). They combined multiple criteria: **alignment to the edit instruction**, **preservation of unedited content (source-target fidelity)**, and **overall visual quality** of the edited image [83]. *Eval type:* **Human MOS (Mean Opinion Score)** – humans rated edited images on those aspects, and scores were z-score normalized for fairness [84]. They also introduced **IE-QA**, a question-answer based automatic check for editing (verifying if the specific edit was done). *Axes:* **Edit success** (did the specific change happen correctly?), **identity preservation** (are other details same as original?), and **image realism/quality**. *Human vs Auto:* Primarily human (with some automatic verification). By providing a dedicated benchmark for editing, this work acknowledges that **image generation and image editing have different evaluation needs** – an edited image must be faithful to an input image *and* to text, a dual constraint general T2I generation doesn't have. Early results show that even if a model is good at generation, it may fail subtle editing tests (e.g. correctly handling shadows or keeping the background identical).

(Other notable 2025 developments): *Multi-objective metrics* emerged to evaluate multiple goals at once. For example, Kim et al. (2025) proposed a method to add a **"reward head"** to a vision-language model that can output scores for various criteria (like relevance, safety, diversity) simultaneously. They even included feedback from *blind or low-vision users* to weight certain criteria (the **EYE4ALL dataset**), highlighting inclusivity in evaluation. There is also growing interest in **open-source evaluation platforms**: the Stanford HELM project integrated T2I evaluation, and community toolkits (like the one for **cFreD**) are being released to let anyone compute these new metrics [85].

Current Evaluation Practices

Today's evaluation of text-to-image models typically **combines multiple approaches** to cover the different axes of quality:

- **Human preference studies** remain the gold standard for subjective aspects. Researchers frequently conduct **pairwise comparisons or rating studies** where human judges choose which image is better for a given prompt, or rate images on Likert scales for criteria like realism, prompt alignment, or aesthetics [86] [87]. For example, OpenAI's DALL·E 2 was partly evaluated by human raters for caption alignment and photorealism, and Stability AI uses human ranking to tune models. Human evaluation captures nuances of perception and preference that automated metrics might miss (e.g. what looks "more correct" to a person).

- **Automatic metrics** are used for scale and objectivity. Common **reference-free metrics** include **CLIPScore** (measuring image–text embedding similarity) and learned reward models like **PickScore** or **ImageReward**. For visual quality without a prompt, metrics like **FID** (lower is better) and **Inception Score** (higher is better) are computed on standard datasets (e.g. MS-COCO) to assess fidelity and diversity. However, **relying on one metric is no longer considered sufficient** – researchers often report a *suite of metrics*. For instance, a model report might include FID for image realism, CLIPScore for overall alignment, **R-Precision** for caption similarity, and **Diversity scores** for variation. If these metrics disagree, it flags the need for further analysis.

- **Multi-dimensional evaluations** are becoming common. Instead of a single "quality" score, evaluations now often break results down by aspects. A recent comprehensive evaluation combined **human ratings** on prompt alignment, photorealism, and detail, *plus* a battery of automated tests: CLIP-based prompt similarity, raw image quality measures (sharpness, contrast, etc.), and even an **LLM-based critique** of each image [88] [89]. In that study, 100 complex prompts were generated by three models and assessed by both humans and an AI (Claude) [90] [91]. The **findings** illustrated why a multi-faceted approach is useful: for example, Stable Diffusion was top-ranked by humans overall, despite having a relatively low technical image quality score (its images were artistically detailed and clear to people, even if somewhat grainy by pixel metrics) [92] [93]. Conversely, one model (Imagen 2) scored high on objective image clarity but humans still preferred others due to better prompt interpretation [94] [95]. Such discrepancies show that *"human perception of quality doesn't always align with traditional metrics"* [93] [95]. Therefore, researchers use **hybrid evaluations** to get a complete picture.

- **Advanced AI evaluators** are now part of the toolkit. With the rise of multimodal large language models, evaluators like **GPT-4V** (vision-enabled GPT-4) or **Claude-Vision** are being used to judge images. For example, one can prompt GPT-4 with: *"Here is a prompt and an image – how well does the image fulfill the prompt? Give a score and rationale."* Studies have found GPT-4's judgments to align impressively well with aggregate human judgments [96] [97]. Leveraging this, some works (e.g. the ACL 2025 T2I-Eval paper) **distill GPT-4's evaluation capability into smaller models** [98] [99] so that an open-source model can rate images across various criteria. This involves training on **GPT-generated ratings** for many prompt–image pairs. The result is an automatic judge that approaches the quality of GPT-4 or human raters, but can be run locally by anyone [100] [99]. While care is needed (models like GPT-4 can have blind spots or unknown biases), this represents a promising shift toward **accessible, AI-based evaluation** that still retains a form of "common-sense" reasoning.

In summary, current best practices use a **mix of human and automated evaluations** to compensate for the weaknesses of each. Humans can assess subtle semantic alignment, compositional coherence, and overall "impression" of an image, whereas automated metrics provide reproducible numbers and can pinpoint specific issues (like KL divergence in feature space, or a missed object via QA). Increasingly, **evaluations are multi-axial** – reporting separate scores for different facets (alignment, realism, diversity, etc.) rather than a single omnibus score. There's also an emphasis on **explainability**: whether via human commentary, modular metrics (like VPEval), or LLM-generated rationales, evaluators try to explain *why* a model got the score it did [32] [101].

Future Evaluation Trends and Open Challenges

Looking ahead, the rapid progress of T2I models calls for **even more rigorous and diverse evaluation**. Here are some **trends and suggestions for the next year (and beyond)**:

- **Structured and** Diagrammatic Outputs**: As the user hinted,** evaluating models on tasks like generating labeled diagrams or infographics **is a next frontier. We can expect benchmarks where the model must produce an image *with certain structures or annotations* (e.g. a flowchart with labeled nodes, a cartoon with speech bubbles, an instructional diagram). These tasks test spatial reasoning and the ability to combine text and graphics. Early work in 2025 has started this with math formulas, code, and charts** [76] **, but going forward, benchmarks might require models to place** text labels on parts of the image **correctly or generate** scene graphs **from prompts and then images from those graphs. Evaluating this could involve using** OCR and object detection **to verify that *each labeled part is present and correctly named*. For example, a future benchmark might say: "Generate an anatomical diagram of a flower labeling the petals, stem, and leaves." Success requires not just rendering a flower, but also placing the words "petals," "stem," "leaves" on the correct parts. To evaluate this, one could parse the generated image with an OCR to read the labels and a detector to locate the parts, checking alignment between them. Structured evaluations like** scene-graph-based scoring **(as done in GenomeBench) could be extended here** [66] [71] **. By** focusing on part-whole relationships and explicit labeling**, we'd test a model's fine-grained understanding of the content it generates, beyond global caption similarity.

- **Long-Form and** Multimodal Understanding**: As prompts become more detailed (think: describing an entire comic strip or a paragraph of a story to illustrate), models will need to handle** long-form coherence**. LPG-Bench has begun addressing this** [79] **, and we will likely see *even longer or multi-turn prompts* in evaluations – e.g. a sequence of prompts that build a story. This bleeds into** text-to-video or sequential image generation**, where consistency between frames (the same characters, etc.) becomes an eval criterion. Even within single images, future benchmarks might include** multi-sentence descriptions with contextual references **(for instance, "The first sentence describes the background, the second describes the foreground; generate an image for the whole paragraph"). Metrics like TIT-Score-LLM, which use an LLM to compare the prompt and a generated caption of the image, will be key for scoring long prompts** [82] **. Researchers should work on** metrics robust to long text**, possibly combining *segment-level evaluations* (check each sentence or clause) and overall semantic similarity. Additionally, as models incorporate** multi-modal inputs *(say, a prompt + a reference image for style or a layout sketch), evaluation must cover how well the model obeyed all* input modalities. For example, if given a reference style image and a text prompt, did the output match the style? This could require new benchmarks (providing different types of conditioning signals) and metrics to compare styles or layouts.

- **Contextual and** Cultural Evaluation**: The cultural context is an important aspect where current models falter** [65] **. We should see** benchmarks for cultural correctness **(e.g. understanding that a prompt about an Indian wedding implies certain attire and rituals, versus a Japanese wedding). This goes hand-in-hand with** fairness and bias **evaluation: designing prompt sets that reveal biases (like how a model portrays different genders or ethnicities for a given occupation) and measuring them. Future work should incorporate** demographic and cultural scenarios **to ensure models' outputs are appropriate and not reinforcing stereotypes. Metrics**

here might remain mostly human-driven (it's hard for an automated metric to know if an image is culturally insensitive), but one can imagine training evaluators (perhaps region-specific) or using approaches like CulturalFrames *to quantify mismatches* [65] *. The goal would be to push models to not only avoid toxic or biased content*, but to *positively respect cultural nuances* in prompts. This could mean multi-lingual and multi-cultural test suites – as hinted by the inclusion of multilinguality and fairness in HEIM [42] – and possibly involving evaluators from the cultures in question for human studies.

- **Combining Multiple Objectives**: Traditionally, we measure one aspect at a time, but real-world usage demands balancing trade-offs: e.g. a user might want **high realism** *and* **no offensive content** *and* **adherence to prompt**. We expect to see evaluation frameworks that **jointly assess multiple objectives**. For instance, a *single metric* might output a weighted score or a vector of scores that include alignment, realism, safety, and diversity. Some recent work (like Multi-TAP by Kim et al.) is moving this way by giving a model an "evaluation head" that outputs scores for various criteria in one go. Next year, researchers should explore **user-customizable evaluation**: e.g. a benchmark or metric that can be tuned to different preferences (one user cares more about creativity, another about accuracy). This could involve collecting *different types of human preference data* (as EYE4ALL did for accessibility) and ensuring our metrics or reward models reflect those diverse preferences. **Composite benchmarks** might also emerge – for example, an evaluation where models must **optimize a blend of scores** (like "the model scores above X on alignment and above Y on realism"). Such multi-objective evaluation will encourage development of models that are well-rounded instead of excelling at one narrow metric.

- **Explainability and** Transparency in Evaluation**: As the evaluation process grows more complex (with LLM judges, ensembles of metrics, etc.), it will be crucial to maintain clarity. One trend will be building** evaluation reports **rather than single scores. We already see this in VPEval providing textual explanations** [32] **, or CLIPScore being augmented with heatmaps of which parts of the prompt were matched. Future work could formalize this: benchmarks might require not just a score but an** explanation **of each decision. For instance, an evaluation system might say: *"Score: 7/10. The image matches the prompt overall, but it missed the 'red striped shirt' detail and the background is wrong."* This mirrors how a human judge would give feedback. Achieving this automatically might involve combining** visual grounding **(to point out where in the image the prompt failed) and** language description **(to articulate it). Such explainable evaluation not only helps model developers debug models, but can itself be a benchmark of the evaluator's quality (does it catch the right issues?). We should encourage research on** evaluators that have interpretable internal reasoning**, possibly using *chain-of-thought prompting* (as some metrics do). In short, the next generation of evaluation should not be a black box; it should** clearly communicate why a model's output is good or bad**, enabling trust in the evaluation process.

- **Robustness and Stress Testing**: Another direction is evaluating how models handle **adversarial or out-of-distribution prompts**. E.g., a benchmark could include deliberately challenging prompts: "A yellow banana on a yellow background" (to test if the model can handle subtle contrast), or nonsense prompts to see if the model over-confidently generates something. Also, testing **consistency under perturbations** – like slightly rewording the prompt or adding irrelevant sentences – and seeing if the evaluation (or the model outputs) change significantly. A robust model should give similar quality output for semantically similar prompts, and a robust metric should not

be fooled by small image perturbations [102] [103]. Work like the *Re-Evaluating Alignment* paper (2025) showed current metrics can be inconsistent under random seeds or tiny image noise [104] [105]. So, future evaluations may include **statistical significance tests** and require models/metrics to maintain rankings across perturbations [106] [107]. Researchers should adopt practices like reporting a **confidence interval or p-value for metric comparisons**, as suggested in that work, to ensure claimed improvements are meaningful [108] [109]. Overall, an emphasis on *evaluation robustness* – not just the mean score, but how stable and reliable it is – will likely grow.

- **Community Benchmarks and Open-Source Tools**: We anticipate more **community-driven benchmarks** similar to how GLUE or SUPERGLUE functioned for NLP. The T2I field might get a centralized leaderboard that evaluates models on a battery of tests (as HEIM did, or a superset of many benchmarks above). There's already mention of things like **GenAI-Bench** and **PartiPrompts Arena**, and integration with evaluation platforms like **LAION's OpenBench** or Stanford's HELM codebase [41] [85]. A trend will be to make these evaluations **accessible and reproducible** – e.g. providing dockerized toolkits where you input a model's API and it spits out scores on multiple benchmarks. The next year could see the rise of a **"benchmark suite"** that includes short prompts, long prompts, compositional tests, safety tests, etc., all together. This would simplify comparing models from different organizations on an even footing. Researchers should contribute to and use these open benchmarks so that evaluation standards converge and improve. The use of **common test sets with public human eval results** (like HEIM's released data [44] or the human annotations in EvalMuse [55]) means we can calibrate new metrics against known human judgment baselines more easily.

In conclusion, evaluating text-to-image models is becoming as complex and multi-faceted as the models themselves. We've moved from simple scores like "FID on COCO" to rich evaluation matrices covering everything from basic object fidelity to abstract qualities like "originality" or "fairness." Researchers in the next year should focus on **completing this evolution**: developing **new benchmarks for new capabilities** (e.g. diagrams, story-length prompts), ensuring **evaluations keep up with model capabilities** (e.g. if a model can generate text in images, our metrics must actually read that text), and emphasizing **transparency and robustness** in the evaluation process. By doing so, we will gain a deeper understanding of our models' strengths and weaknesses, ultimately driving the field toward models that not only produce beautiful images but also reliably do what we intend, in all the ways we care about [110] [111].

Sources:

- Multi-task T2I benchmark (2022) – Petsiuk et al. NeurIPS'22 [2] [3]
- TISE metrics suite (2022) – Dinh et al. ECCV'22 [6] [8]
- DrawBench & prompt eval – Saharia et al. (Imagen 2022); Yu et al. (Parti 2022) [10] [12]
- Pick-a-Pic/PickScore (2023) – Kirstain et al. NeurIPS'23 [18] [20]
- TIFA metric and benchmark (2023) – Hu et al. ICCV'23 [26] [28]
- ImageReward model (2023) – Xu et al. NeurIPS'23 [31] [112]
- VPEval framework (2023) – Cho et al. NeurIPS'23 [32]
- Ross et al. "Good Metric?" analysis (2024) – COLM'24 [58] [37]
- Holistic Eval (HEIM) (2023) – Lee et al. NeurIPS'23 [42] [41]
- HPS v2 and HPD dataset (2024) – Wu et al. arXiv'24 [46] [51]
- EvalAlign (2024) – Tan et al. 2024
- PN-VQA and EvalMuse (2024) – Han et al. 2024 [54] [55]

- Re-Evaluating alignment (2025) – Nie et al. 2025 [105] [108]
- cFreD metric (2025) – Koo et al. CVPR'25 [63]
- CulturalFrames (2025) – Nayak et al. CVPR'25 [65]
- GenomeBench (2025) – Corneanu et al. WACV'25 [67] [73]
- Specialized text benchmark (2025) – Bosheah & Bilicki, MDPI'25 [76] [77]
- LPG-Bench & TIT-Score (2025) – Wang et al. 2025 [79] [82]
- IE-Bench for editing (2025) – Sun et al. 2025 [113]
- Labelbox eval study (2023) – Labelbox Inc. [89] [92]
- ACL 2025 T2I-Eval (2025) – Tu et al. ACL'25 [96] [98]

---

[1] [2] [3] [4] cs.columbia.edu
https://www.cs.columbia.edu/~idrori/Human_Evaluation_of_Text_to_Image_Models_NeurIPS_HEGM_2022.pdf

[5] [6] [7] [8] ecva.net
https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136960585.pdf

[9] [10] [11] [12] [23] [24] [25] [26] [27] [28] TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering
https://openaccess.thecvf.com/content/ICCV2023/papers/Hu_TIFA_Accurate_and_Interpretable_Text-to-Image_Faithfulness_Evaluation_with_Question_Answering_ICCV_2023_paper.pdf

[13] [14] [15] [16] [17] [18] [19] [20] [21] [22] proceedings.neurips.cc
https://proceedings.neurips.cc/paper_files/paper/2023/file/73aacd8b3b05b4b503d58310b523553c-Paper-Conference.pdf

[29] [112] zai-org/ImageReward - GitHub
https://github.com/zai-org/ImageReward

[30] [PDF] ImageReward: Learning and Evaluating Human Preference for Text ...
https://neurips.cc/media/neurips-2023/Slides/72054.pdf

[31] [PDF] ImageReward: Learning and Evaluating Human Preferences for ...
https://papers.nips.cc/paper_files/paper/2023/file/33646ef0ed554145eab65f6250fab0c9-Paper-Conference.pdf

[32] [33] [34] [101] Visual Programming for Step-by-Step Text-to-Image Generation and Evaluation | OpenReview
https://openreview.net/forum?id=yhBFG9Y85R&referrer=%5Bthe%20profile%20of%20Abhay%20Zala%5D(%2Fprofile%3Fid%3D~Abhay_Zala1)

[35] Davidsonian Scene Graph: Improving Reliability in Fine-grained ...
https://arxiv.org/html/2310.18235v4

[36] What makes a good metric? Evaluating automatic metrics for text-to ...
https://www.researchgate.net/publication/387183650_What_makes_a_good_metric_Evaluating_automatic_metrics_for_text-to-image_consistency

[37] [38] [58] [59] [60] [61] openreview.net
https://openreview.net/pdf?id=LFfktMPAci

[39] [40] [41] [42] [43] [44] [2311.04287] Holistic Evaluation of Text-To-Image Models
https://arxiv.org/abs/2311.04287

45 46 47 48 49 50 51 52 53 [2306.09341] Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis
https://ar5iv.labs.arxiv.org/html/2306.09341

54 55 56 57 62 63 64 65 82 83 84 85 110 111 113 Text-Image Alignment Metrics: A Review
https://www.emergentmind.com/topics/text-image-alignment-metrics

66 67 68 69 70 71 72 73 74 Structured Human Assessment of Text-to-Image Generative Models
https://openaccess.thecvf.com/content/WACV2025/papers/Corneanu_Structured_Human_Assessment_of_Text-to-Image_Generative_Models_WACV_2025_paper.pdf

75 76 77 78 Challenges in Generating Accurate Text in Images: A Benchmark for Text-to-Image Models on Specialized Content
https://www.mdpi.com/2076-3417/15/5/2274

79 LPG-Bench: Long-Prompt Benchmark for T2I Models - Emergent Mind
https://www.emergentmind.com/topics/lpg-bench

80 81 TIT-Score: Evaluating Long-Prompt Based Text-to-Image Alignment ...
https://arxiv.org/html/2510.02987v1

86 87 88 89 90 91 92 93 94 95 A comprehensive approach to evaluating text-to-image models
https://labelbox.com/guides/a-comprehensive-approach-to-evaluating-text-to-image-models/

96 97 98 99 100 aclanthology.org
https://aclanthology.org/2025.acl-long.1088.pdf

102 103 104 105 106 107 108 109 Re-Thinking the Automatic Evaluation of Image-Text Alignment in Text-to-Image Models
https://arxiv.org/html/2506.08480v1