DynEval: Holistic Evaluations of T2I Generative Models in the Wild

Shyam Marjit[1*], Dheeraj Baiju[1*], Anuj Shikarkhane[1*†], Akhil Sakthieswaran[1†], Sayak Paul[2], and Anirban Chakraborty[1]

[1]Indian Institute of Science [2]Hugging Face
shyammarjit@iisc.ac.in
Project Page: https://vcl-iisc.github.io/dyneval
Dataset: https://huggingface.co/datasets/vcl-iisc/DynEval-dataset

Abstract. Recent advances in text-to-image (T2I) generation have led to models capable of producing highly realistic images. Yet, reliably evaluating their outputs remains challenging, especially at scale. Existing automatic evaluators, often relying on a static prompt set, struggle to capture subtle failure modes such as partial prompt misalignment, compositional errors or visually plausible but semantically incorrect generations. In this work, we introduce **DynEval**, a **Dyn**amic **Eval**uation framework designed to jointly assess *text-to-image alignment* and *image quality* of T2I models. To support scalable training beyond limited human-annotated data, we construct two large datasets. First, we build **GenDB**, a collection of 500K prompt-image pairs generated from human-written prompts drawn from DiffusionDB using a tiered prompt-model generation strategy. Second, building upon GenDB, we construct **DynEvalInstruct**, a 250K instruction dataset comprising prompt-image-response triplets distilled from a structured evaluation pipeline that decomposes evaluation into text-image alignment and visual quality reasoning. Using this dataset, we perform full fine-tuning of a compact evaluator through a curriculum learning strategy to effectively distill the superior evaluation capabilities of a larger teacher vision-language model, resulting in **DynEval-2B** and **DynEval-4B**. In extensive comparisons against existing evaluators across 11 benchmarks, our evaluator achieves a higher overall correlation with human judgments. Furthermore, it provides fine-grained analysis of the capabilities and failure modes of **36** T2I models across 42 subcategories and 9 semantic dimensions.

1 Introduction

Text-to-image (T2I) generation has evolved from multi-stage U-Net [47] based diffusion models such as SDXL [44] to diffusion transformers [43] that scale efficiently to high fidelity outputs [17, 32, 33]. In parallel, efficiency-focused diffusion designs have reduced training and deployment costs while enabling high-

* Equal contribution.
† Work done during internship at VCL, IISc.

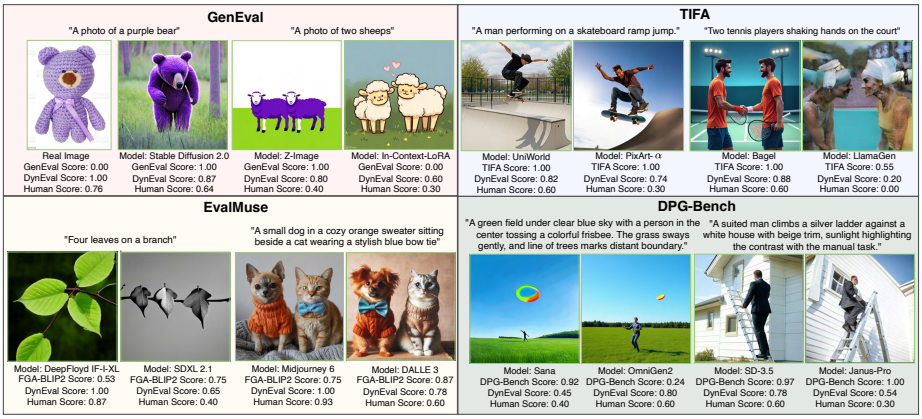

Fig. 1: Qualitative comparison of **DynEval-4B** with four representative text-to-image evaluation methods [20,21,26,27] across multiple T2I models [5,9,11,16,17,29,40,44,46, 50,52,62,64]. All scores are normalized to the [0, 1] range. The first GenEval example additionally includes a real reference image, illustrating the limitations of detector-based evaluation. Compared with existing methods, **DynEval-4B** produces scores that better agree with human judgments by jointly evaluating text–image alignment and image quality. The examples illustrate three representative failure modes of existing evaluators: **(i)** Detector-based methods such as GenEval [20] often fail to accurately evaluate generated content when the outputs exhibit artistic styles or visual distributions that are underrepresented or absent in the detectors' training data; **(ii)** visually superior images may receive disproportionately low scores; and **(iii)** visually distorted or semantically implausible images may receive overly high scores. Additional qualitative comparisons are provided in Supp. Fig. S1.

resolution synthesis on modest hardware [8–10,64,65]. Alongside diffusion, autoregressive generation has emerged as a competing paradigm for visual synthesis [50, 58]. Most recently, unified multi-modal and agentic systems increasingly combine generation with understanding and editing, spanning decoupled-encoder designs, mixture-of-experts reasoning, and hybrid pipelines that emphasize strong text rendering and alignment [6,11,16,19,49,57]. With this rapid advancement, evaluating the fine-grained quality and semantic faithfulness of generated content has become a central challenge, as generated images still exhibit T2I misalignment and visual errors (semantic infeasibility, distortion, *etc*). While human evaluation remains the gold standard for assessing T2I models, it is inherently subjective, expensive, and difficult to scale. Consequently, these limitations have motivated the development of automatic evaluation metrics that aim to mimic human judgment, while remaining scalable.

Early evaluation metrics, including distribution-based measures [3, 23, 48], CLIP [45]-based metrics [22, 42], and captioning-based variants [1, 14, 24, 25, 54] offer an incomplete judgement of model performance as they evaluate a limited subset of dimensions and fail to capture many nuanced aspects of image generation. To address this, several structured evaluation methods emerged.

Detector-based approaches [20] verify object-centric attributes but remain sensitive to distortions, texture artifacts, and unrealistic scenes. To further improve semantic assessment, subsequent works adopted visual question answering (VQA)-based evaluation frameworks [13, 26, 27, 61]. However, they depend on proprietary LLMs for decomposing prompts into targeted question-answer (QA) generation and open-source VQA models for scoring, thereby making evaluation costly and difficult to reproduce at scale. Furthermore, approaches that integrate multiple task-specific modules for evaluation [28] inherit error accumulation due to the integration of different models. To mitigate such error propagation, works like [38, 59, 61] rely entirely on closed source models as evaluators, inherently making evaluation at scale resource-intensive. In contrast, VQAScore [41] highlights the utility of open-source vision language models (VLM) as alternatives to proprietary models using the model's probability that an image matches the prompt. However, this formulation can still miss fine-grained distortions and localized compositional failures. To address this limitation, GenEval2 [30] extends VQAScore with Soft-TIFA [27] scoring, but still relies on external LLM generated questions. Interestingly, all of the aforementioned methods lack an end-to-end trained evaluator for generated prompt-image pairs.

Recent works like EvalMuse [21] and LMM4LMM [55] focus on tuning the scoring VLM using limited ($\approx$ 40-50K) human-annotated ratings on structured questions. While being effective, this human-dependence raises the fundamental question ① . Additionally, existing evaluators overlook fine-grained image quality assessment, often producing inconsistent or contradictory results, where low-quality generations receive higher scores while visually superior images are ranked lower (see Fig. 1). These observations highlight the need for ② .

① *Can we train a small-scale VLM-as-a-judge for T2I model evaluation without relying on human ratings, considering the high cost and limited scalability of large-scale human labeling?*
② *Can we build a more reliable and truly dynamic evaluation framework that aligns closely with human judgment, captures real-world perceptual and semantic quality, and adapts dynamically to open-set prompts without relying on external LLMs at inference time?*

In this work, we move towards these two goals by introducing **DynEval**, a **Dyn**amic **Eval**uator for T2I models, obtained through full fine-tuning of a small-scale model using a curriculum learning strategy that effectively distills the evaluation capabilities of a larger VLM in a structured manner. To support this training, we construct two large-scale datasets. First, we curate **GenDB**, a 500K unique ⟨prompt, image⟩ pair dataset, where images are generated from human-written prompts sampled from DiffusionDB [60]. Instead of uniformly sampling prompts for image generation, GenDB employs a *tier-matched generation strategy*: prompts are grouped by complexity, T2I models are grouped by capability, and prompts are assigned to models according to their corresponding tiers, ensuring that more capable models are exposed to more challenging prompts. This

Table 1: Comparison of capabilities across major T2I-evaluation methods. Existing methods differ widely in their coverage of key evaluation dimensions such as alignment, image quality, object distortions, realism, and dynamic reasoning. Notably, most methods rely on fixed prompt sets and need costly human annotations for all data points, thereby limiting scalability. In contrast, **DynEval** introduces a flexible, scalable, and truly dynamic evaluation framework that is independent of a static prompt set. Here, ◐, ●, and ○ denote partial, full, and no support for a capability, respectively.

Capability	**GENEVAL [20]** (NeurIPS 23)	**TIFA [27]** (ICCV 23)	**DPG-Bench [26]** (DSG [13]-based)	**T2I-CompBench++ [28]** (TPAMI 25)	**EvalMuse-40K [21]** (AAAI-26)	**DynEVAL** (Ours-26)
Text-to-Image Alignment	●	●	●	●	●	●
Image Quality Assessment	○	○	○	◐	◐	●
Dynamic Evaluation	○	◐	◐	○	◐	●
Object Distortions Identification	○	○	○	○	◐	●
Realism Checking	○	○	○	○	◐	●
3D-Spatial Relationships	○	○	○	●	○	◐
Fine-tuned Judge	○	○	○	○	●	●
Tuning Dataset Size	○	○	○	○	40K	250K
Human Annotation for Tuning	○	○	○	○	●	○

strategy improves the coverage of informative failure modes across varying levels of prompt complexity and T2I model capability. From the curated prompt-image pairs in GenDB, we construct **DynEvalInstruct**, a 250K ⟨prompt, image, response⟩ triplets dataset using the teacher model `Qwen3-VL-235B`, enabling large-scale supervision through knowledge distillation beyond the scale achievable with manually annotated datasets. Using the DynEvalInstruct dataset, we train compact **DynEval-2B** and **DynEval-4B** models as automated evaluators that jointly assess text-to-image alignment (T2IA) and image quality (IQA) for every prompt-image pair. To the best of our knowledge, we are the first to introduce compact evaluators to perform both assessments within a unified framework. For T2IA, DynEval generates prompt-grounded verification questions to determine whether the generated image satisfies the semantic requirements of the prompt. For IQA, it constructs a scene graph from the generated image using the text prompt as contextual guidance, and subsequently leverages this graph to generate image-specific evaluation questions targeting realism, structural consistency, and fine-grained object-level failures.

The key contributions of this work are summarized, as follows:

- We construct **GenDB**, a large-scale prompt-image dataset with well-balanced prompt coverage and image generations from **36** T2I models, and derive **DynEvalInstruct** from it, an instruction-tuning dataset to train evaluators at scale beyond limited human-annotated data (refer to Fig. 2 and Sec. 3-4).
- Unlike static QA methods, we propose **DynEval**, a dynamic evaluator that jointly evaluates prompt-image alignment as well as builds a scene graph from generated image to compose structured, image-specific questions for fine-grained image quality assessment (refer to Tab. 1 and Fig. 3).
- To obtain a robust evaluator, we introduce tier-based prompt categorization with tier-specific image generation to cover a wide gamut of failure modes across varying prompt complexities and model capabilities (refer to Fig. 2).
- We conduct the most comprehensive benchmarking of T2I evaluators to date across **11** benchmark datasets. Our evaluator achieves higher overall agree-

ment with human judgments (refer to Tab. 2-3), while also providing fine-grained analyses of the capabilities and failure modes of **36** T2I models across **42** subcategories spanning **9** semantic dimensions (refer to Fig. 4).

2 Related Work

Metrics for evaluating T2I models. Traditional metrics such as FID [23], KID [3], and IS [48] quantified realism and diversity by comparing the feature distributions of generated and reference images using a pretrained Inception-V3 [51] model, while LPIPS [69] captured perceptual similarity. However, these image-only measures rely on reference images and assume that class-based features can represent visual realism, making them unsuitable for open-domain, text-conditioned generation [18]. To address T2I alignment, earlier methods used cosine similarity between DINO [7] image embeddings and CLIP [45] image–text embeddings (e.g., CLIPScore [22] and CLIP-R [42]), while captioning-based approaches converted generated images into textual descriptions [14, 24, 25] and compared them to prompts using language metrics such as CIDEr [54] and SPICE [1]. Recent frameworks [20, 26–28] further emphasize compositional fidelity through object verification and QA-based evaluation.

Object detection based benchmarks. T2I evaluation methods often use object detection (e.g., `Mask2Former` [12]) to measure compositional fidelity by verifying whether generated images accurately depict prompt elements. Early works such as SOA [24] and DALL-Eval [14] check basic properties including object presence, count, and color, while GenEval [20] further combines object detection with CLIP [45] to evaluate entities, attributes, co-occurrence, and spatial relationships. However, these methods remain limited by the coverage and reliability of the underlying detectors and are generally less suited for abstract prompts, subtle distortions, global realism, or image-quality failures.

VQA-based benchmarks. VQA-based evaluation methods leverage LLMs to decompose prompts into QA pairs, enabling more fine-grained and interpretable faithfulness assessment than CLIP-based metrics. TIFA [27] uses `GPT-3` [4] to generate verifiable QA pairs covering attributes such as color, count, and spatial relations, which are evaluated using `mPLUG` [36] or `BLIP-2` [37]. DPG-Bench [26] extends this paradigm via DSG [13], decomposing complex prompts into atomic predicates verified by `mPLUG` [36]. More recently, TIIF-Bench [61] evaluates T2I models under varying prompt lengths. GenEval2 [30] addresses benchmark drift in GenEval by replacing binary object-level scoring with Soft-TIFA, a VQAScore-style [41] primitive-level evaluator designed to better track human judgments. T2I-Eval-Bench [53] decomposes holistic image evaluation into simpler sub-tasks and distills the resulting capability into an open-source MLLM. Similarly, T2I-CoReBench [38] separates explicit composition from implicit reasoning by evaluating scene-graph elements and probing deductive, inductive, and abductive inference using `Gemini 2.5 Flash`. UniGenBench++ [59] broadens semantic evaluation through a hierarchical taxonomy of themes and evaluation criteria, while also testing robustness across English/Chinese and short/long prompt variants

using `Gemini 2.5 Flash`. These works demonstrate that question-based decomposition enables fine-grained diagnosis; however, most prior VQA-style protocols primarily focus on prompt satisfaction via discrete semantic checks, often overlooking image realism (*i.e.* IQA), structural distortions, and perceptual quality.

Hybrid benchmarks. These approaches combine VQA, object detection, and traditional metrics to provide a more holistic assessment of T2I alignment [28,56]. T2I CompBench++ [28] introduces specialized metrics, including disentangled BLIP-VQA for attribute binding, UniDet [71]-based evaluation for 2D/3D spatial relationships, and a fine-tuning scheme to strengthen prompt alignment. Similarly, OmniGenBench [56] adopts a dual-mode design: perception-centric evaluation uses visual parsing and scene analysis tools, while cognition-centric evaluation uses `Gemini-2.5-Pro` [15] as a judge to assess consistency, realism, and aesthetics. Complementing these, HEIM [34] combines human evaluation and traditional metrics across 12 dimensions, including alignment, aesthetics, bias, and robustness.

Human preference guided benchmarks. Beyond objective alignment and fidelity metrics, prior works incorporate human preference to capture higher-level qualities such as aesthetics, creativity, and perceptual appeal. ImageReward [67] trains a reward model from human ratings and pairwise comparisons to rank images according to human preference for alignment and fidelity. VisionReward [66] learns an interpretable human-preference score by decomposing visual quality into hierarchical dimensions and fine-grained binary checks, followed by weighted regression on human annotations and pairwise preferences. EvalMi-50K [55] and EvalMuse-40K [21] construct human-annotated evaluation datasets spanning perception and text–image correspondence to train the underlying judge model. In contrast, our framework trains the evaluator entirely without human annotations.

3 *GenDB* Dataset

Overview. Our goal is to develop a dynamic and robust evaluator that diagnoses T2I model failure modes under *real-world user prompts*. To achieve this, we begin by constructing **GenDB**, a large-scale dataset designed to capture diverse prompt complexities and model capabilities. We then curate prompts using a complexity-based scoring rule (*prompt tiering*) and stratify 36 T2I models by capability using the **DynEval-1K** evaluation set (*model tiering*). Finally, we generate 500K ⟨prompt, image⟩ pairs through tier-matched prompt-model assignment. GenDB serves as the foundation for constructing **DynEvalInstruct** (Sec. 4), which is used to train the **DynEval** models (Sec. 5).

Complexity-based Scoring. We sample prompts from **DiffusionDB** [60], which contains 1.8M unique prompts specified by real users, to better reflect real-world user prompting while avoiding the handcrafted templates and LLM-generated prompts commonly used in existing benchmarks. From this large pool, we curate challenging prompts using a filtering and ranking pipeline. Concretely,

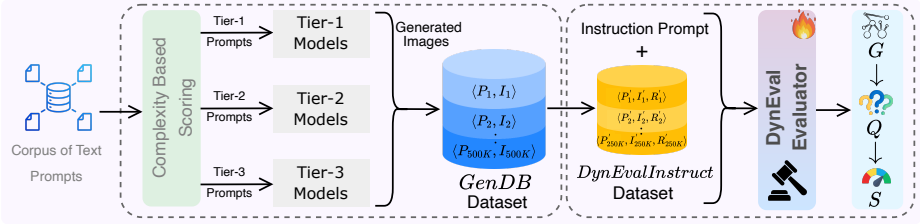

Fig. 2: Overview of the DynEval Evaluation Framework. GenDB consists of a diverse set of ⟨prompt, image⟩ pairs (noted as $\langle P_i, I_i \rangle$) obtained by categorizing prompts and models into three tiers (refer to Sec. 3), where one image is generated per prompt. DynEval fine-tunes a VLM on **DynEvalInstruct** dataset (construction detailed in Sec. 4) to generate structured evaluation outputs consisting of T2IA and IQA. T2IA generates prompt-grounded verification questions and answers to assess semantic alignment. IQA is composed of a scene graph (G) inferred jointly from the image and prompt and a set of detailed questions (Q) for each node and relation in the graph. Over all verification questions generated by the T2IA and IQA processes, we perform VQA-based evaluation and obtain a score S between 1 and 5, assigned to each question.

each prompt is assigned a **heuristic complexity score** based on **9** factors including *prompt length*, *object and attribute counts*, and other 7 *semantic attributes*. We discard short prompts and rank the remainder by complexity score. More details on the considered factors and how these factors translate to the complexity score are provided in the Supp. Sec. C.

Prompt Tiering. We construct a pool of 500K prompts from DiffusionDB [60] by selecting the diverse and challenging prompts using the aforementioned *complexity based scoring* rule. After ranking by complexity, we split the selected 500K prompts into three difficulty tiers using two adaptive score thresholds (τ_1 and τ_2), yielding Tier-1 (hard, score $> \tau_1$), Tier-2 (medium, $\tau_2 <$ score $\leq \tau_1$), and Tier-3 (easy, score $\leq \tau_2$) prompts. Details of the thresholding strategy and tier-wise illustrative prompt examples are provided in the Supp. Sec. D.

Model Tiering. Now, to group the 36 T2I models into three capability tiers, we carefully sample 1,000 prompts from the evaluation prompt pools of [20, 21, 26, 27], ensuring uniform coverage across 42 subcategories spanning 9 semantic dimensions (*e.g.*, 'Object & Entity' covers subcategories like *Human* and *Vehicle*; see Sec. 6.3). Thereafter, each of these 1000 prompts is processed by all 36 T2I models, resulting in 1,000 images per model and a total of 36K ⟨prompt, image⟩ pairs. We refer to this dataset as the **DynEval-1K** evaluation set, which is also later used for model ablations and failure analysis. Now, each pair from this set of 36K is scored using `Qwen3-VL-235B` [2], the same large VLM used as the teacher model during distillation. Following our holistic evaluation framework (Fig. 3), the score aggregates two aspects: T2IA and IQA (same strategy as described in detail in Sec. 4). We then average scores across the 1,000 prompts to obtain a final score for each model. Finally, we apply two thresholds ($\mu_1 > \mu_2$) to these

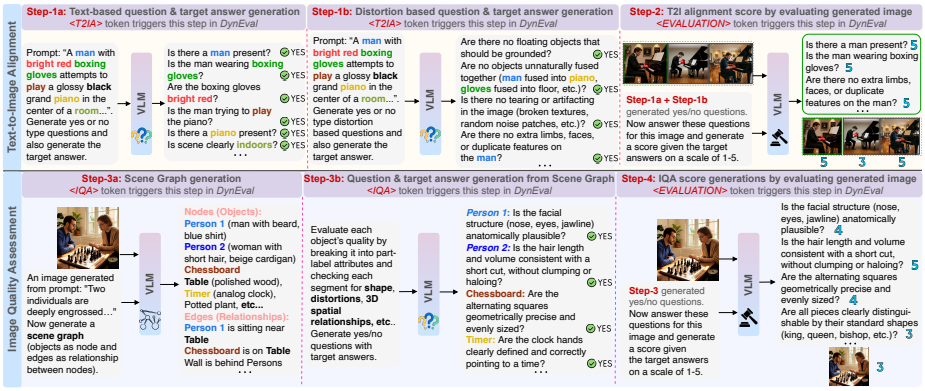

Fig. 3: Overview of the proposed DynEvalInstruct generation pipeline and DynEval training framework. The teacher VLM (`Qwen3-VL-235B` [2]) decomposes evaluation into two complementary dimensions: **T2IA**, which generates prompt-grounded semantic and distortion-checking questions and target answers, and **IQA**, which constructs scene graphs and object-level quality assessment questions and answers. The teacher model then performs VQA-based scoring on the generated question sets to obtain T2IA and IQA assessments. The resulting dataset consists of ⟨prompt, image, response⟩ triplets, where the **response** comprises the complete structured T2IA and IQA evaluation outputs generated in Steps 1–4, including generated questions, target answers, scene graphs, and question-level evaluation scores. These annotations collectively provide fine-grained supervision for training **DynEval** with three task-specific tokens: `<T2IA>`, `<IQA>`, and `<EVALUATION>`.

aggregated scores to partition models into three tiers, where Tier-1 denotes the strongest group and Tier-3 comprises the weakest among the considered models. The exhaustive list of tier assignments is provided in the Supp. Sec. E.

Tier-matched pairing to construct GenDB. Finally, once the T2I models and prompts are grouped into three tiers, we construct **GenDB**, a dataset of 500,000 ⟨prompt, image⟩ pairs. Specifically, we use the previously sampled 500K prompts (from DiffusionDB [60]) and generate images using 36 T2I models while enforcing tier consistency, i.e., Tier-i models are paired exclusively with Tier-i prompts. To maximize prompt diversity under a fixed computational budget, we do not generate images from all 36 models for each prompt. Instead, within each tier, prompts are randomly partitioned into disjoint subsets and assigned to individual models, ensuring that each prompt is processed by exactly one model. This tiered structure is motivated by the observation that Tier-3 models often perform poorly even on simple prompts, making evaluation on more complex prompts less effective. Conversely, Tier-1 models tend to perform well on simple prompts, yielding fewer failure cases for training a robust evaluator. Therefore, it is necessary to pair models with prompts commensurate with their category.

4 *DynEvalInstruct* Dataset

Overview. While constructing GenDB, we employ complexity-based prompt sorting and prompt-model tier matching to increase the likelihood of capturing failure cases in T2I generation. However, we observe that many ⟨prompt, image⟩ pairs still exhibit reasonably good prompt-image alignment, making them less informative for training a robust evaluator which requires exposure to diverse and fine-grained failure modes that capture subtle semantic, compositional, and visual inconsistencies. To address this limitation, we introduce a second-stage curation framework that selectively identifies informative prompt-image pairs exhibiting diverse failure modes. A large teacher VLM (`Qwen3-VL-235B` [2]) is employed to perform structured evaluation of each ⟨prompt, image⟩ pair along two complementary dimensions: Text-to-Image Alignment (T2IA) and Image Quality Assessment (IQA). These evaluations are subsequently used for two purposes: (a) dataset curation, which when applied to GenDB yields **DynEvalInstruct**, a large-scale instruction dataset, and (b) providing fine-grained supervision for training a lightweight text-to-image evaluator. Now, we describe the structured evaluation framework used to obtain these annotations, followed by the curation strategy used to construct DynEvalInstruct.

Structured Evaluation of a ⟨prompt, image⟩ pair. Given a ⟨prompt, image⟩ pair, we obtain T2IA and IQA scores through a structured intermediate process of QA generation followed by VQA-based scoring, as illustrated in Fig. 3. This pipeline also serves as a knowledge distillation framework, transferring the evaluation capabilities of a large-scale teacher VLM to our lightweight evaluator. By decomposing evaluation into T2IA and IQA responses, we distill fine-grained supervision from the teacher model rather than relying on a single overall score. We next describe the QA generation and VQA-based scoring process in detail.

(i) Text-to-Image Alignment (T2IA). In this stage, we evaluate how well a generated image aligns with its corresponding text prompt.

- **QA Generation:** Given a text prompt, the teacher VLM generates atomic binary 'yes/no' questions to verify prompt alignment across objects, attributes, actions, spatial relations, scene consistency, *etc.* It also generates complementary *distortion-focused* questions targeting likely visual failures associated with the prompt (*e.g.*, object fusion, floating objects, broken textures, and implausible structures). These are combined into a unified T2IA question set with corresponding target answers (**Steps 1a-1b** in Fig. 3).
- **VQA-based Scoring:** The combined question set and the image generated from the same prompt are then passed to the teacher VLM which assigns a score in [1, 5] for each question based on the given input rubric. The question-wise scores are averaged to obtain overall alignment score (**Step-2** in Fig. 3).

(ii) Image Quality Assessment (IQA). While T2IA measures prompt faithfulness, the IQA branch evaluates the intrinsic visual quality of the generated image. Its goal is to identify whether the objects and relations present in the image are visually plausible, structurally coherent, and free from local artifacts or

distortions. The text prompt is used only as contextual reference during scene-graph construction, while the questions themselves are grounded in the visual content of the generated image.

- **Scene Graph Generation:** Given an image, the VLM generates a scene graph with nodes representing objects and edges denoting relationships between them. To constrain generation, we provide text prompt as a reference input, thereby implicitly reducing the likelihood of hallucinating objects and relationships not present in the image, compared to image-only VLM input counterparts (**Step-3a** in Fig. 3).
- **Image Quality QA Generation:** For each node in the scene graph, the VLM decomposes the corresponding object into its constituent attributes and generates a set of binary 'yes/no' questions targeting quality dimensions such as shape consistency, texture fidelity, completeness, structural realism, and 3D spatial cues (*e.g.,* depth perception and relative distance). These questions, along with each of their corresponding 'yes' or 'no' answers, form the IQA question-answer set, enabling a detailed evaluation of object-wise and attribute-wise quality for each generated image (**Step-3b** in Fig. 3).
- **VQA-based Scoring:** We use the same scoring methodology as the VQA-based scoring in the T2IA pipeline, extending it to IQA questions for IQA scoring (**Step-4** in Fig. 3).

(iii) Combining T2IA and IQA Scores. For a given ⟨prompt, image⟩ pair, we first compute the T2IA and IQA scores and combine them via a weighted sum: $\alpha \times$ (T2IA score) $+ \beta \times$ (IQA score), where both α and β are set to 0.5.

Curation of DynEvalInstruct. Building upon the structured evaluation framework described above, we run the teacher VLM on each prompt-image pair in GenDB, yielding 500K $\langle \text{prompt}, \text{image}, \text{response} \rangle$ triplets, where each response contains structured T2IA and IQA evaluation outputs produced by the teacher model (for details refer to Fig. 3). The resulting T2IA and IQA scores are then weighted-averaged to obtain a single scalar score S. Since, GenDB images originate from three model tiers with inherently different performance ranges, we use **tier-specific thresholds** δ_i (for $i \in \{1, 2, 3\}$). For a sample generated by a Tier-i model, we mark it as a *failure case* if $S < \delta_i$, and add it to a selected set $\mathcal{D}$. This yields a filtered subset of 250K triplets ($|\mathcal{D}| = 250\text{K}$), referred to as the **DynEvalInstruct** dataset. Notably, empirical analysis reveals that performance saturates at around 250K training samples, with diminishing returns beyond this point; accordingly, we construct DynEvalInstruct using 250K selected triplets (see Supp. Tab. S11). The resulting dataset maintains balanced representation across prompt tiers while ensuring coverage of challenging semantic, compositional, and visual failure cases across the full spectrum of model capabilities. Further details on the thresholds are provided in the Supp. Sec. F.

5 DynEVAL

Training an evaluator for T2I models at scale requires reliable data which is infeasible to obtain through human annotation for hundreds of thousands of

Table 2: Zero-shot Quantitative comparison between DynEval and existing methods for predicting overall alignment scores across multiple benchmarks. Columns correspond to different datasets containing ⟨prompt, image⟩ pairs with associated human ratings, while rows represent evaluation methods used to estimate alignment scores. Notably, **DynEval** evaluators are trained entirely without human annotations yet achieve strong agreement with human judgments across diverse evaluation settings. GenEval [20] scores are reported only for annotation-compatible datasets; unavailable EvalMuse [21] entries are left blank. Best: ●, Second-best: ●.

Evaluation Method	EvalMuse [21]		GenAI-B [41]		TIFA [27]		RichHF [39]		GenEval [20]		EvalMI [55]		T2I-Eval-B [53]	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
CLIPScore [22]	0.2993	0.2933	0.1676	0.2030	0.3003	0.3086	0.0570	0.3024	0.0654	0.0985	0.2607	0.3072	0.0678	0.1260
BLIPv2Score [37]	0.3583	0.3348	0.2734	0.2979	0.4287	0.4543	0.1425	0.3105	0.2669	0.3167	0.2900	0.3468	0.1223	0.1321
ImageReward [67]	0.4655	0.4585	0.3400	0.3786	0.6211	0.6336	0.2747	0.3291	0.4149	0.4603	0.4991	0.5523	0.1349	0.1144
PickScore [31]	0.4399	0.4328	0.3541	0.3631	0.4279	0.4342	0.3916	0.4133	0.2787	0.3015	0.4611	0.4692	0.2375	0.2440
HPSv2 [63]	0.3745	0.3657	0.1371	0.1693	0.3647	0.3804	0.1871	0.2577	0.3223	0.3580	0.5336	0.5525	0.2916	0.2990
VQAScore [41]	0.4877	0.4841	0.5534	0.5175	0.6951	0.6585	0.4826	0.4094	0.5057	0.5040	0.6062	0.6118	0.3348	0.3570
GenEval [20]	-	-	-	-	-	-	-	-	0.4753	0.4509	-	-	-	-
FGA-BLIP2-OS [21]	**0.7742**	**0.7722**	**0.5637**	**0.5673**	**0.7604**	**0.7442**	0.5123	0.5455	0.5170	0.5278	0.6724	0.6945	0.3582	0.3471
TIIF-Bench [61]	-	-	0.4387	0.4020	0.6007	0.5776	0.4342	0.4292	0.5600	0.5447	0.4474	0.4256	0.3297	0.3016
T2I-Eval-Bench [53]	-	-	0.4996	0.4773	0.6011	0.6246	0.5033	0.4990	0.5847	0.5610	0.4258	0.4075	0.3744	0.3629
UniGenBench++ [59]	-	-	0.4605	0.4441	0.6078	0.5892	0.5333	0.5172	0.5661	0.5494	0.4611	0.4554	0.3971	0.3818
T2I-CoReB (8B) [38]	-	-	0.4614	0.4414	0.6126	0.6112	**0.6024**	**0.5809**	0.5626	0.5810	0.5583	0.5414	0.4035	0.3879
GenEval 2 (8B) [30]	-	-	0.4817	0.4571	0.6824	0.6734	**0.6311**	**0.6021**	**0.6044**	**0.6175**	**0.7245**	**0.7098**	0.4222	0.4172
LongT2IBench [68]	-	-	0.3747	0.3763	0.6775	0.6922	0.5174	0.5161	0.4544	0.5999	0.5056	0.5485	**0.4781**	**0.4563**
DynEval-2B (Ours)	0.7765	0.7798	0.5715	0.5695	0.7827	0.7715	0.5226	0.5545	0.5668	0.5820	0.7346	0.7156	0.4535	0.4410
DynEval-4B (Ours)	**0.7932**	**0.7945**	**0.5945**	**0.5817**	**0.8022**	**0.8034**	0.5457	0.5691	**0.5894**	**0.6073**	**0.7944**	**0.7889**	**0.4857**	**0.4636**

⟨prompt, image⟩ pairs. We therefore adopt a distillation-based approach, where the capabilities of a strong teacher VLM are transferred to smaller evaluator models. We propose **DynEval**, with two variants: DynEval-2B and DynEval-4B, trained via distilling responses from `Qwen3-VL-235B` [2]. A key design choice is that we do not distill DynEval as a black-box regressor that directly maps a ⟨prompt, image⟩ pair to a single score. Importantly, VLMs tend to produce redundant questions when directly asked to generate questions conditioned on either text prompt (**T2IA** process) or the image (**IQA** process), which violates atomicity by repeatedly verifying overlapping attributes and relations, thereby artificially inflating the final evaluation score. Motivated by [13, 21], we also emphasize **atomicity** and **coverage**: questions should be atomic, avoid hallucinations, and maintain valid dependencies. To implement this structured distillation, we introduce three **task-specific tokens** that explicitly trigger different evaluation procedures during training and inference (refer to Fig. 3).

- `<T2IA>` triggers generations of diverse binary verification questions using only the prompt. With this token, the model also generates distortion-checking questions with a separate instruction-tuning prompt.
- `<IQA>` activates scene-graph generation and fine-grained image-quality question generation with respective instruction-tuning prompt.
- `<EVALUATION>` triggers VQA-style answering and scoring. Given a generated image and a set of questions produced by `<T2IA>` or `<IQA>` token, the model answers each question and outputs a score in the range [1-5].

These objectives are trained in a curriculum: the evaluator first learns to generate structured questions and then learns to answer them. This reduces model collapse during fine-tuning and greatly stabilizes the evaluator's reasoning behavior. In practice, we find that the tokenized curriculum helps the model disentangle *semantic alignment*, *visual integrity*, and *relative preference*–three crucial dimensions that previous evaluators either conflated or ignored.

6 Results and Discussions

6.1 Experimental Setup

Training Settings. We do full-scale fine-tuning of `Qwen3-VL-4B` to obtain **DynEval-4B** and `Qwen3-VL-2B` to obtain **DynEval-2B**. The models are trained using the **DynEvalInstruct** dataset consisting of 250K samples, where each sample contains an image, a text prompt, and a detailed response generated through the T2IA and IQA pipelines (refer to Fig. 3). For training, we use a linear warmup followed by cosine decay, with a peak learning rate of 2×10^{-5}. The models are trained for one epoch on 4x NVIDIA H200 GPUs.

Evaluation Settings. We compare DynEval with **14** existing evaluation methods on **11** benchmarks by reporting the Spearman Rank Correlation Coefficient (**SRCC**) and Pearson Linear Correlation Coefficient (**PLCC**) to measure the correlation between the evaluator predicted score and human-annotated score. Higher correlation values indicate stronger agreement with human judgment. We use results reported in the original evaluator papers whenever available; otherwise, we evaluate official checkpoints on newer challenging benchmarks. Tab. 2 reports comparative results on the **7** benchmarks that offer publicly available prompts, generated images, and human annotations, while the additional benchmarks in Tab. 3 lack human annotations. Detailed statistics of the considered evaluation benchmarks are provided in Supp. Sec. A.

6.2 Experimental Results.

Qualitative Results. In Fig. 1, we showcase comparison between prediction scores of **DynEval-4B** with existing evaluation methods [20,21,26,27] on representative image-text pairs. As shown, popular T2I evaluations, *e.g.,* GenEval [20] and DPG-Bench [26], often assign high scores to distorted images despite their low human ratings. Similarly, methods like EvalMuse [21] and TIFA [27] also exhibit substantial deviations from human judgments. In contrast, our proposed DynEval produces scores that are consistently more aligned with human evaluation, demonstrating its effectiveness in jointly assessing image-text alignment and image quality. Further qualitative results are shown in Supp. Fig. S1.

Quantitative Results. As shown in Tab. 2-3, **DynEval-4B** achieves SOTA performance on 9 out of 11 benchmarks and ranks second best on one additional benchmark. This demonstrates strong generalization across diverse text-to-image evaluation settings. Averaged over all benchmarks in Tab. 2-3, DynEval-4B surpasses the previous best evaluators by a relative margin of +**4.77%**, with the relative improvement increasing to +**7.61%** on the benchmarks where it establishes a new state of the art in terms of SRCC. Notably, DynEval-4B achieves the largest relative gains on EvalMI [55] (+9.65%) and GenAI-Bench [35] (+5.46%) in Tab. 2, while attaining substantial relative improvements on UniGenBench++ [59] (+20.87%) and TIIF-Bench [61] (+8.83%) in Tab. 3. Furthermore, on the RichHF and GenEval benchmarks, DynEval-4B trails the larger 8B evaluators T2I-CoReBench [38] and GenEval2 [30] by relative margins of 9.41% and 2.48%, respectively, while only using half the parameters of these methods. Nonetheless, DynEval-4B achieves the strongest overall performance across the complete suite of evaluation benchmarks. These results highlight the effectiveness and robustness of DynEval as a unified and highly generalizable T2I model evaluation framework.

Table 3: Additional zero-shot comparison between DynEval and recent evaluation methods. Columns denote benchmark datasets and rows denote evaluators. Best: ●; second-best: ●.

Evaluation Method	GenEval2 [30]		TIIF-B [61]		UniGenB++ [59]		T2I-CoReB [38]	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
VQAScore [41]	0.6223	0.5208	0.3728	0.3353	0.3109	0.2991	0.2469	0.2668
FGA-BLIP2-OS [21]	0.3868	0.3750	0.3581	0.4437	0.3521	0.3457	0.3154	0.3075
TIIF-Bench [61]	0.3432	0.3322	0.3917	0.3690	0.3254	0.3111	0.3602	0.3562
T2I-Eval-Bench [53]	0.3904	0.3862	0.3392	0.3190	0.3473	0.3356	0.3811	0.3767
UniGenBench++ [59]	0.4602	0.4072	0.4788	0.4553	0.3888	0.3752	0.3671	0.3454
T2I-CoReB (8B) [38]	0.4813	0.4955	0.5196	0.5671	0.4004	0.3849	**0.4469**	**0.4352**
GenEval2 (8B) [30]	**0.7120**	**0.7017**	**0.5495**	**0.6028**	**0.4102**	**0.3858**	0.4435	0.4315
DynEval-4B (Ours)	**0.7508**	**0.7132**	**0.5980**	**0.6298**	**0.4958**	**0.5042**	**0.4856**	**0.4716**

Model and Data Scaling. Furthermore, scaling our evaluator from 2B to 4B parameters consistently improves performance across all benchmarks in Tab. 2, yielding an average relative SRCC gain of **4.62%**. The largest improvements occur on EvalMI (+**8.14%**) and T2I-Eval-Bench (+**7.10%**), while even benchmarks with smaller gains such as EvalMuse (+**2.15%**) show consistent improvement. This demonstrates that DynEval benefits from model scaling. We further provide ablations on teacher model selection (Tab. S10) and fine-tuning dataset scaling (Tab. S11), analogous to model scaling, in the Supp.

Statistics on Dynamic Question Generation. In Tab. 4, we validate the dynamic IQA question generation process by analyzing the number and quality of questions generated by **DynEval-4B** across 11 evaluation datasets. It is observed that the number of

Table 4: Statistics and quality analysis of **DynEval-4B** generated questions to assess IQA across 11 evaluation datasets, using `Qwen3-VL-235B` as Teacher.

# Objects in Image	Question Count Statistics				Quality of IQA Generated Questions	
	Mean	Median	Min	Max	Coverage Ratio ↑	BERTScore ↑
1–5	14.2	14	5	25	0.885	0.899
6–10	32.2	33	24	40	0.824	0.871
11–15	47.6	45	36	61	0.752	0.817
16–20	66.6	63	47	100	0.696	0.772

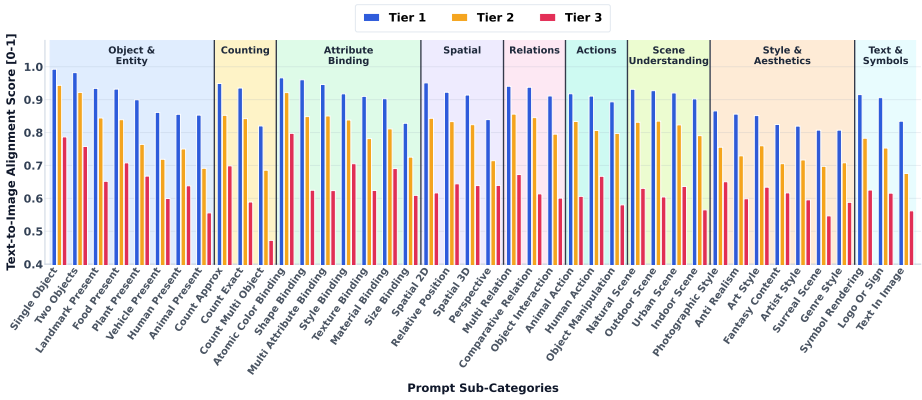

Fig. 4: Fine-grained capability analysis of 36 T2I models on the DynEval-1K evaluation set using DynEval-4B across 42 prompt sub-categories and 9 semantic dimensions. While Tier-1 models exhibit stronger alignment across most categories, all tiers show substantial performance degradation on challenging aspects such as *count multi object*, *style and aesthetics*, *human present*, and *size binding*. These findings indicate that, despite significant progress in object grounding and attribute binding, these capabilities remain persistent challenges for current T2I models.

questions scales almost linearly with number of objects in the image. In addition to reporting question-count statistics, we also evaluate the quality of the generated IQA questions independently of the final DynEval score. Specifically, using `Qwen3-VL-235B` [2] as the teacher VLM, we compute: **(i) Coverage Ratio**, defined as the IoU between teacher and student response sets; and **(ii) BERTScore similarity** [70], to measure semantic alignment between teacher-generated and DynEval-generated questions, both reported in the range ([0-1]).

6.3 Understanding T2I-Models' Failure Attributes

As shown in Fig. 4, while modern T2I models demonstrate strong capabilities in basic object grounding, scene understanding, and standard spatial and reasoning, several challenges persist across all capability tiers, which can be summarized as follows: **(i)** Exact object counting emerges as one of the most challenging capabilities, with performance degrading rapidly as compositional complexity increases by number of objects and multi-attribute objects. **(ii)** Prompts involving humans and animals continue to be significantly more difficult than those involving inanimate entities (*e.g.,* landmarks, food, and vehicles), highlighting challenges in modeling fine-grained structures and appearances. **(iii)** Similarly, models struggle with complex attribute binding, particularly under counterfactual or uncommon size relationships that contradict real-world priors, suggesting a strong reliance on learned visual priors rather than robust relational reasoning. **(iv)** Perspective-dependent prompts involving diverse camera viewpoints remain challenging even for the most capable models, highlighting limitations in

geometric understanding and viewpoint control. **(v)** Accurate text and symbol generation also remains an unresolved challenge, with models frequently failing to faithfully preserve textual content within generated images. Although higher-capability models achieve better overall alignment, the relative difficulty ordering of these challenging dimensions remains largely consistent across model tiers. To further analyze this behavior, we provide comparisons between the best and worst models within each tier across all 42 evaluation subcategories in the Supp. Sec. I.

Despite rapid advances in text-to-image (T2I) generation, reliable evaluation remains an open challenge. As illustrated in Fig. 5, recent T2I models achieve steadily improving scores on benchmarks such as GenEval [20]. However, as a detector-based framework, GenEval mainly rewards object and attribute presence, such that performance gains may reflect improved attribute satisfaction rather than genuine advances in visual fidelity. Although recent VQA-based evaluators improve semantic assessment, they remain largely alignment-centric and overlook fine-grained perceptual quality [21]. In contrast, our approach jointly evaluates text-image alignment and image quality, yielding a more holistic assessment and substantially stronger agreement with human judgments.

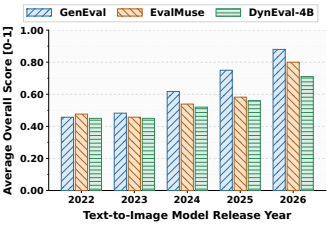

Fig. 5: Average scores assigned by existing methods [20, 21] and DynEval-4B to T2I models grouped by publication year. While existing evaluators show steadily increasing scores for newer models, DynEval-4B produces more calibrated assessments by jointly evaluating text-image alignment and image quality (visual realism).

7 Conclusion

In this work, we presented **DynEval**, a holistic and truly dynamic evaluation framework for text-to-image models that jointly assesses *text-to-image alignment* (T2IA) and *image quality* (IQA), targeting failure modes that are often missed or inconsistently scored by existing automated evaluators. To enable scalable training without manual annotations, we constructed **GenDB**, a 500K prompt-image dataset, and subsequently derived from it **DynEvalInstruct**, a 250K instruction dataset distilled from a teacher VLM. Leveraging this data, we trained lightweight evaluators (DynEval-2B/4B) that achieve overall state-of-the-art performance across **11** benchmarks while also providing fine-grained diagnostic insights into persistent weaknesses of modern T2I models.

Acknowledgements

We gratefully acknowledge Kotak-IISc AI/ML Centre (KIAC) for the generous conference travel grant and the GPU resources that helped this research.

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: European conference on computer vision. pp. 382–398. Springer (2016)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
5. Cai, H., Cao, S., Du, R., Gao, P., Hoi, S., Hou, Z., Huang, S., Jiang, D., Jin, X., Li, L., et al.: Z-image: An efficient image generation foundation model with single-stream diffusion transformer. arXiv preprint arXiv:2511.22699 (2025)
6. Cao, S., Chen, H., Chen, P., Cheng, Y., Cui, Y., Deng, X., Dong, Y., Gong, K., Gu, T., Gu, X., et al.: Hunyuanimage 3.0 technical report. arXiv preprint arXiv:2509.23951 (2025)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
8. Chen, J., Xue, S., Zhao, Y., Yu, J., Paul, S., Chen, J., Cai, H., Han, S., Xie, E.: Sana-sprint: One-step diffusion with continuous-time consistency distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16185–16195 (2025)
9. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: International conference on learning representations. vol. 2024, pp. 57611–57640 (2024)
10. Chen, J., Zhao, Y., Yu, J., Chu, R., Chen, J., Yang, S., Wang, X., Pan, Y., Zhou, D., Ling, H., et al.: Sana-video: Efficient video generation with block linear diffusion transformer. arXiv preprint arXiv:2509.24695 (2025)
11. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
12. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
13. Cho, J., Hu, Y., Baldridge, J., Garg, R., Anderson, P., Krishna, R., Bansal, M., Pont-Tuset, J., Wang, S.: Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In: International conference on learning representations. vol. 2024, pp. 15625–15645 (2024)
14. Cho, J., Zala, A., Bansal, M.: Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3043–3054 (2023)
15. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)

16. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
17. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Proceedings of the 41st International Conference on Machine Learning. pp. 12606–12633 (2024)
18. Frolov, S., Hinz, T., Raue, F., Hees, J., Dengel, A.: Adversarial text-to-image synthesis: A review. Neural Networks **144**, 187–209 (2021)
19. Geng, Z., Wang, Y., Ma, Y., Li, C., Rao, Y., Gu, S., Zhong, Z., Lu, Q., Hu, H., Zhang, X., et al.: X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. arXiv preprint arXiv:2507.22058 (2025)
20. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
21. Han, S., Fan, H., Fu, J., Li, L., Li, T., Cui, J., Wang, Y., Tai, Y., Sun, J., Guo, C.L., et al.: Evalmuse-40k: A fine-grained benchmark with comprehensive human annotations for text-to-image generation model alignment evaluation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 4583–4591 (2026)
22. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
23. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
24. Hinz, T., Heinrich, S., Wermter, S.: Semantic object accuracy for generative text-to-image synthesis. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1552–1565 (2020)
25. Hong, S., Yang, D., Choi, J., Lee, H.: Inferring semantic layout for hierarchical text-to-image synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7986–7994 (2018)
26. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 (2024)
27. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20406–20417 (2023)
28. Huang, K., Duan, C., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(5), 3563–3579 (2025)
29. Huang, L., Wang, W., Wu, Z.F., Shi, Y., Dou, H., Liang, C., Feng, Y., Liu, Y., Zhou, J.: In-context lora for diffusion transformers. arXiv preprint arXiv:2410.23775 (2024)
30. Kamath, A., Chang, K.W., Krishna, R., Zettlemoyer, L., Hu, Y., Ghazvininejad, M.: Geneval 2: Addressing benchmark drift in text-to-image evaluation. arXiv preprint arXiv:2512.16853 (2025)
31. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in neural information processing systems **36**, 36652–36663 (2023)

32. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024)
33. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
34. Lee, T., Yasunaga, M., Meng, C., Mai, Y., Park, J.S., Gupta, A., Zhang, Y., Narayanan, D., Teufel, H., Bellagente, M., et al.: Holistic evaluation of text-to-image models. Advances in Neural Information Processing Systems **36**, 69981–70011 (2023)
35. Li, B., Lin, Z., Pathak, D., Li, J., Fei, Y., Wu, K., Xia, X., Zhang, P., Neubig, G., Ramanan, D.: Evaluating and improving compositional text-to-visual generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5290–5301 (2024)
36. Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., et al.: mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In: Proceedings of the 2022 conference on empirical methods in natural language processing. pp. 7241–7259 (2022)
37. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
38. Li, O., Wang, Y., Hu, X., Huang, H., Chen, R., Ou, J., Tao, X., Wan, P., Qi, X., Feng, F.: Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? arXiv preprint arXiv:2509.03516 (2025)
39. Liang, Y., He, J., Li, G., Li, P., Klimovskiy, A., Carolan, N., Sun, J., Pont-Tuset, J., Young, S., Yang, F., et al.: Rich human feedback for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19401–19411 (2024)
40. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147 (2025)
41. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: European Conference on Computer Vision. pp. 366–384. Springer (2024)
42. Park, D.H., Azadi, S., Liu, X., Darrell, T., Rohrbach, A.: Benchmark for compositional text-to-image synthesis. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021)
43. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
44. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: International Conference on Learning Representations. vol. 2024, pp. 1862–1874 (2024)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
46. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

47. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
48. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016)
49. Song, Y., Dong, W., Wang, S., Zhang, Q., Xue, S., Yuan, T., Yang, H., Feng, H., Zhou, H., Xiao, X., et al.: Query-kontext: An unified multimodal model for image generation and editing. arXiv preprint arXiv:2509.26641 (2025)
50. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
51. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
52. Team, D.: Deepfloyd if: A modular cascaded diffusion model for text-to-image synthesis. `https://github.com/deep-floyd/IF` (2023), stability AI Research
53. Tu, R.C., Ma, Z.A., Lan, T., Zhao, Y., Huang, H.Y., Mao, X.L.: Automatic evaluation for text-to-image generation: Task-decomposed framework, distilled training, and meta-evaluation benchmark. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 22340–22361 (2025)
54. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
55. Wang, J., Duan, H., Zhao, Y., Wang, J., Zhai, G., Min, X.: Lmm4lmm: Benchmarking and evaluating large-multimodal image generation with lmms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17312–17323 (2025)
56. Wang, J., Jiao, Y., Yu, Y., Qian, T., Chen, S., Chen, J., Jiang, Y.G.: Omnigenbench: A benchmark for omnipotent multimodal generation across 50+ tasks. arXiv preprint arXiv:2505.18775 (2025)
57. Wang, P., Peng, Y., Gan, Y., Hu, L., Xie, T., Wang, X., Wei, Y., Tang, C., Zhu, B., Li, C., et al.: Skywork unipic: Unified autoregressive modeling for visual understanding and generation. arXiv preprint arXiv:2508.03320 (2025)
58. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
59. Wang, Y., Li, Z., Zang, Y., Bu, J., Zhou, Y., Xin, Y., He, J., Wang, C., Lu, Q., Jin, C., et al.: Unigenbench++: A unified semantic evaluation benchmark for text-to-image generation. arXiv preprint arXiv:2510.18701 (2025)
60. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In: Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers). pp. 893–911 (2023)
61. Wei, X., Zhang, J., Wang, Z., Wei, H., Guo, Z., Zhang, L.: Tiif-bench: How does your t2i model follow your instructions? arXiv preprint arXiv:2506.02161 (2025)
62. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)

63. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
64. Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., et al.: Sana: Efficient high-resolution image synthesis with linear diffusion transformers. arXiv preprint arXiv:2410.10629 (2024)
65. Xie, E., Chen, J., Zhao, Y., Yu, J., Zhu, L., Wu, C., Lin, Y., Zhang, Z., Li, M., Chen, J., et al.: Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. arXiv preprint arXiv:2501.18427 (2025)
66. Xu, J., Huang, Y., Cheng, J., Yang, Y., Xu, J., Wang, Y., Duan, W., Yang, S., Jin, Q., Li, S., et al.: Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 11269–11277 (2026)
67. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 15903–15935 (2023)
68. Yang, Z., Gu, T., Wang, J., Lin, F., Sheng, X., Chen, P., Li, L.: Longt2ibench: A benchmark for evaluating long text-to-image generation with graph-structured annotations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 11820–11828 (2026)
69. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
70. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
71. Zhou, X., Koltun, V., Krähenbühl, P.: Simple multi-dataset detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7571–7580 (2022)