DSH-Bench: A Difficulty- and Scenario-Aware Benchmark with Hierarchical Subject Taxonomy for Subject-Driven Text-to-Image Generation

Zhenyu Hu[1⋆], Qing Wang[1⋆], Te Cao[1⋆], Luo Liao[1⋆], Longfei Lu[1], Liqun Liu[1†], Shuang Li[1], Hang Chen[1], Mengge Xue[1‡], Yuan Chen[1], Chao Deng[1], Peng Shu[1], Huan Yu[1], and Jie Jiang[1]

Tencent, China
{mapleshu, rosaliecao, graceqwang, berryxue, liqunliu}@tencent.com

Abstract. Significant progress has been achieved in subject-driven text-to-image (T2I) generation, which aims to synthesize new images depicting target subjects according to user instructions. However, evaluating these models remains a significant challenge. Existing benchmarks exhibit critical limitations: 1) insufficient diversity and comprehensiveness in subject images, 2) inadequate granularity in assessing model performance across different subject difficulty levels and prompt scenarios, and 3) a profound lack of actionable insights and diagnostic guidance for subsequent model refinement. To address these limitations, we propose DSH-Bench, a comprehensive benchmark that enables systematic multi-perspective analysis of subject-driven T2I models through four principal innovations: 1) a hierarchical taxonomy sampling mechanism ensuring comprehensive subject representation across 58 fine-grained categories, 2) an innovative classification scheme categorizing both subject difficulty level and prompt scenario for granular capability assessment, 3) a novel Subject Identity Consistency Score (SICS) metric demonstrating a 9.4% higher correlation with human evaluation compared to existing measures in quantifying subject preservation, and 4) a comprehensive set of diagnostic insights derived from the benchmark, offering critical guidance for optimizing future model training paradigms and data construction strategies. Through an extensive empirical evaluation of 19 leading models, DSH-Bench uncovers previously obscured limitations in current approaches, establishing concrete directions for future research and development.

Keywords: Single Subject-driven · Text-to-image Generation · Dataset and Benchmark · Subject Identity Consistency

⋆ Equal contribution. † Corresponding author. ‡ Project leader.

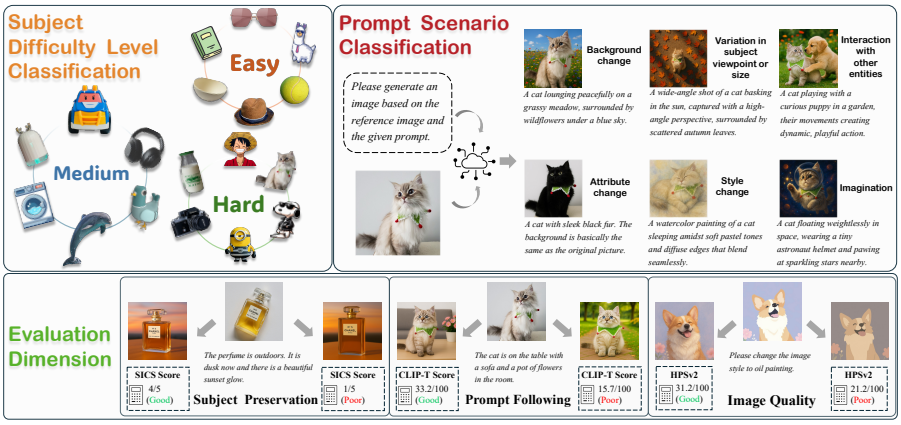


Fig. 1: Overview of DSH-Bench. We curate a diverse dataset of subject images and categorize them into three difficulty levels—**easy**, **medium**, and **hard**—based on the complexity of preserving subject details. Leveraging GPT-4o's capabilities, we systematically generate contextually appropriate prompts for various scenarios. The generated images are then rigorously evaluated across three key dimensions: **Subject Preservation**, **Prompt Following**, and **Image Quality**.

1 Introduction

Subject-driven text-to-image (T2I) generation, which synthesizes novel scenes conditioned on specific reference images and textual prompts, has emerged as a pivotal research frontier, propelled by rapid advancements in large-scale T2I diffusion models [3, 6, 10, 11, 14, 28, 51, 55]. In subject-driven T2I generation, aside from image quality, two other fundamental criteria must be satisfied: Subject Preservation and Prompt Following. Subject Preservation requires that the generated image maintain the details of the reference subject. Prompt Following demands that the generated image consistently reflects the content in the prompt. For example, a user might request an image of "his dog traveling around the world". In this scenario, the generated image must depict a dog identical to the reference image while illustrating the act of traveling as described.

Although significant progress has been made in subject-driven T2I generation in recent years [15, 17, 23, 31, 33, 49, 54, 63, 67, 74], the community still lacks a standardized protocol to comprehensively and effectively evaluate their true capabilities. An optimal benchmark should not only ensure unbiased, human-aligned evaluation but also provide granular diagnostics to guide research. However, current benchmarks [7, 31, 46, 54, 64] are limited by insufficient diversity and comprehensiveness in subject image collection, which restricts the thoroughness of model evaluation. Crucially, they fail to disentangle the inherent difficulty of the reference subjects from the complexity of the prompt scenarios. As shown in Fig. 2, our empirical analysis reveals a significant performance variance contingent on these factors: models that effortlessly reconstruct simple geometries (e.g., a

tennis ball) frequently fail to preserve the intricate structural details of complex artifacts (e.g., a camera). This observation highlights a critical blind spot in current evaluations: the necessity of stratifying test cases by the subject difficulty and prompt scenario. Moreover, subject-driven T2I generation can be broadly categorized into single-subject and multi-subject paradigms depending on the number of reference subjects. While the single-subject task can be conceptually considered a special case of the multi-subject scenario, the two paradigms impose fundamentally different demands on model capabilities. Regarding the scope of evaluation, we posit that the single-subject paradigm is the cornerstone of subject-driven generation. While multi-subject generation introduces interactional complexities, empirical evidence suggests that models still struggle to achieve saturated performance even in single-subject tasks. Consequently, establishing a rigorous benchmark for the single-subject scenario is a prerequisite for any reliable multi-subject assessment.

To address the aforementioned limitations, we introduce DSH-Bench, a novel and comprehensive benchmark designed to provide multi-dimensional evaluations. An overview of DSH-Bench is illustrated in Fig. 1. DSH-Bench offers three distinct advantages:

1. ***The diversity of subject images in DSH-Bench is substantially greater*** To effectively mitigate potential evaluation bias stemming from limited subject diversity, we construct a rigorous hierarchical taxonomy for image collection. Specifically, we incorporate established ontologies from COCO [35] and ImageNet [9] into our taxonomy design. As illustrated in Fig. 3(a), the widely utilized DreamBench comprises merely 6 categories and 30 subjects. In sharp contrast, our benchmark significantly scales up the dataset to 58 distinct categories and 459 unique subjects—representing a substantial increase of **9**× and **15**×, respectively. While DreamBench++ [46] offers 150 subjects, its diversity remains constrained by a limited collection scope. Notably, **33%** of our categories are entirely absent from the DreamBench++ distribution. Consequently, benefiting from DSH-Bench's superior subject diversity, we facilitate a far more comprehensive and robust evaluation of generative models.

2. ***An innovative classification scheme for subject difficulty and prompt scenarios*** As illustrated in Fig. 2, model performance exhibits significant variance across different samples, underscoring the critical necessity for a dual classification of both subject images and prompts. While DreamBench++ attempts to categorize prompts based on perceived difficulty, the underlying criteria remain ambiguous and lack a systematic definition. Furthermore, it neglects to analyze the inherent difficulty levels associated with distinct subjects. To address these limitations, we systematically stratify subjects into three difficulty tiers (easy, medium, hard) based on the complexity of preserving visual appearance, and classify prompts into six distinct scenarios (background change, variation in subject viewpoint or size, interaction with other entities, attribute change, style transfer, and imagination). Thus, our approach enables a more comprehensive and granular diagnosis of the challenges faced by current models.

3. ***A human-aligned and more efficient metric for subject preservation*** DreamBench++ replaces CLIP [50] and DINO [5] with GPT-4o [42] for evaluation, resulting in improved alignment with human evaluation. However, our benchmark reveals that per-model evaluation under this paradigm requires approximately 20,000 API calls to GPT-4o, incurring prohibitive computational costs exceeding $400 for each evaluation. To address the limitation, we introduce **Subject Identity Consistency Score** (SICS), which innovatively focuses on subject-level consistency rather than merely relying on embedding comparisons. Firstly, five annotators label a training dataset containing 5,000 image-text pairs, focusing on subject preservation evaluation. We then fine-tune Qwen2.5-VL-7B [2] on this dataset, which leads the model to focus on core visual attributes rather than high-level semantics. Finally, we use Kendall's τ value to quantify the alignment between model outputs and human evaluation. Experimental results demonstrate that SICS achieves a statistically significant improvement, outperforming Dreambench++ by 9.4% in human evaluation correlation metrics.

Takeaways We present some insightful findings from evaluating 19 methods: i) Our evaluation reveals that no single method demonstrates consistently robust performance across all categories. Therefore, implementing hierarchical taxonomy sampling of subject images is critical for mitigating potential evaluation biases. ii) All methods exhibit degraded performance on hard subject images. It is crucial to enhance models' ability to encode and reconstruct complex subject details more effectively in future research. iii) The subject-driven T2I model's capability for different prompt scenarios is not robust. Future research on subject-driven T2I generation should focus on optimizing for adaptation to a variety of prompt scenarios.

In summary, our contributions are as follows: 1) We employ a hierarchical taxonomy in image collection to ensure both the diversity and comprehensiveness of subject images. 2) We propose an innovative classification scheme to categorize subject difficulty levels and prompt scenarios. This scheme enables us to obtain valuable insights. 3) We propose a human-aligned metric to evaluate subject preservation, which offers greater efficiency compared to DreamBench++. We open-source DSH-Bench, including subject images, prompts and code, at `https://github.com/sunriverhzy/DSH-Bench`.

2 Related Work

2.1 Subject-Driven Text-to-Image Generation

In recent years, subject-driven T2I generation has attracted significant research attention [15–17, 23, 31, 33, 49, 54, 63, 67]. Within the context of diffusion models, optimization-based model [20, 25, 37, 62] enables subject-driven generation by introducing lightweight parameters and performs parameter-efficient fine-tuning for each subject. In contrast, the encoder-based methods [8, 21, 22, 26, 27, 32, 34, 38, 43, 53, 56, 68, 71, 75] leverage additional image encoders and network layers to encode the reference image of the subject. IP-Adapter [74] introduces cross-attention through an additional image encoder to

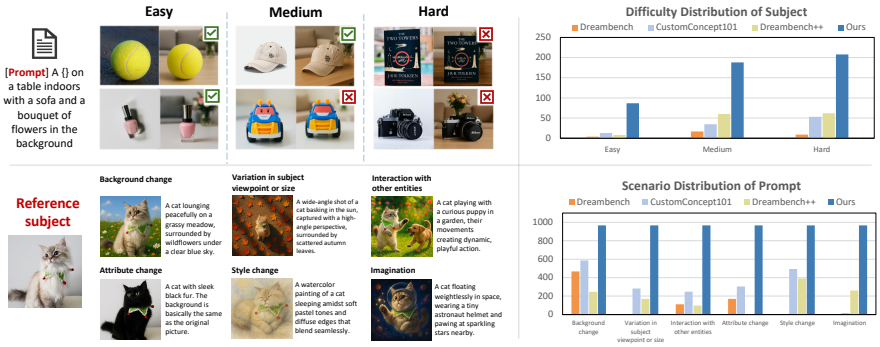
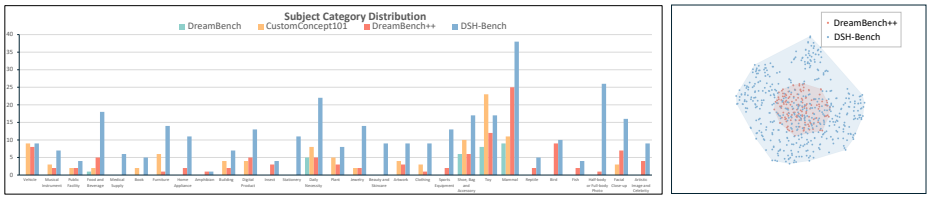


Fig. 2: Qualitative comparison under different difficulty levels and scenarios.



(a) Distribution of Subject Images Across Different Categories (Under Photorealistic Category)

(b) t-SNE Visualization of DreamBench++ vs. DSH-Bench

Fig. 3: Distribution of subject images. (a) Category-wise image distribution for our benchmark versus prior benchmarks. (b) t-SNE comparison of images between DSH-Bench and DreamBench++.

incorporate control signals. Furthermore, SSR-Encoder [77] enhances identity preservation without necessitating further fine-tuning when introducing new concepts. The Diffusion Transformers [45, 47, 52] uses transformer as a denoising network to iteratively refine noisy image tokens. Based on these foundation models, approaches like OminiControl [60] and UNO [68] explore the inherent image reference capabilities of transformers.

2.2 Subject-Driven T2I Generation Benchmark

Evaluation for subject-driven T2I generation involves a variety of metrics focusing on different aspects. For image quality, several notable studies [1, 30, 65, 69, 73] have conducted. For subject preservation evaluation, learning-based metrics [12, 48, 76] compute the distances between image features extracted by deep neural networks. Image embeddings from large vision models like CLIP [50], DINO [5] and image-retrieval score [36] have been utilized. To better match human perception, DreamSim [13] focuses on foreground objects when evaluating image similarity. In terms of semantic consistency, the CLIP score is frequently used. More recently, RefVNLI [57] jointly evaluates textual alignment and subject preservation within a single prediction. Dreambench [54] is the first benchmark established for subject-

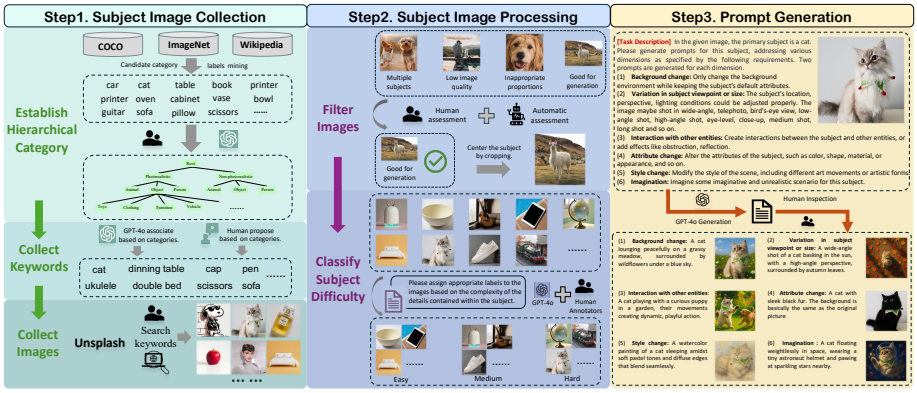


Fig. 4: Dataset construction process of DSH-Bench. We construct a hierarchical taxonomy to obtain a comprehensive set of keywords. Then we collect web images using these keywords. After performing both manual review and automated filtering of the images, we classify the difficulty of subject images and use GPT-4o to generate prompts for each subject image.

driven T2I tasks. However, Dreambench is limited in the diversity of subjects and prompt scenarios. DreamBench-v2 [7] extends the evaluation by introducing 220 novel prompts. CustomConcept101 [31] lacks of non-photorealistic subject images. Although DreamBench++ [46] increases to 150 subject images, it can not provide a systematic categorization of subjects and prompts, which makes it difficult to derive meaningful insights from the evaluation results. More recent efforts such as MMIG-Bench [24] further broaden evaluation toward multi-modal image generation across diverse tasks. To bridge this gap, we introduce DSH-Bench, which enables a more comprehensive evaluation of models for subject-driven T2I generation and provides deeper insights.

3 DSH-Bench

This section provides an overview of the primary components of DSH-Bench. Sec. 3.1 outlines the data construction process. Sec. 3.2 introduces the definitions and evaluation methods for three evaluation dimensions. *A detailed explanation is available in the supplementary materials.*

3.1 Benchmark Dataset Construction

Subject Image Collection ***1) Hierarchical Taxonomy Establishment*** As illustrated in Fig. 4, we construct a rigorous three-tier taxonomy. The first level distinguishes between *Photorealistic* and *Non-photorealistic* domains, with strictly aligned subcategories to ensure cross-domain consistency. For the second level, we refine the coarse "Living" category from prior benchmarks [46, 54] into

distinct *Humans* and *Animals*, acknowledging the unique demand for facial fidelity in human subjects versus the high variance in animals. We explicitly exclude abstract "Style" categories [31] to focus strictly on tangible entity customization. At the third level, to balance granularity with generality, we compiled candidate labels from COCO and ImageNet, utilizing GPT-4o to consolidate them into 58 fine-grained categories. Notably, we stratify the *Human* category into celebrities, facial close-ups, and full-body shots, allowing us to disentangle foundation model bias (e.g., celebrity overfitting) from structural reconstruction capabilities. ***2) Keyword Collection & Internet Image Collection*** In DreamBench++, keywords collection relies on GPT-4o and human input. The approach does not adequately ensure the diversity of the obtained keywords, potentially introducing bias during the image collection process. In contrast, DSH-Bench derives keywords from a hierarchical taxonomy. For each third-level category, we use GPT-4o to generate associated keywords, which are further supplemented by humans. All keywords are then consolidated and deduplicated, resulting in a final set of **400** unique keywords—surpassing DreamBench++'s 300. The specific keywords are provided in the Supplementary material (Sec. B). Given a set of selected keywords, we retrieve images from Unsplash [61]. *Each image's copyright status has been verified for academic suitability.*

Subject Image Processing ***1) Image Filtering*** To filter unsuitable images, we use aesthetic score [72] and SAM [29] to filter images with low image quality and inappropriate proportions of subject regions. The curated images are subsequently cropped to centralize the reference subject. ***2) Subject Difficulty Level Classification*** As illustrated in Fig. 2, the model's performance varies considerably across different samples. To derive meaningful insights, we classify the subject images according to the difficulty level that the model experiences in preserving details of the reference subject. We define three subject difficulty levels, including (1) **Easy:** Subjects characterized by minimal surface complexity and homogeneous textural properties, exemplified by smooth-surfaced objects such as a ceramic mug with uniform coloration. These cases present negligible challenges for detail preservation due to their structural regularity. (2) **Medium:** Subjects containing discernible high-frequency features while maintaining global structural coherence, such as cylindrical containers with legible typographic elements. These cases require intermediate detail preservation capabilities. (3) **Hard:** Subjects exhibiting non-uniform texture distributions and multi-scale geometric details, typified by objects like book covers containing fine-grained calligraphic elements. Such cases expose model limitations in maintaining structural fidelity and textural granularity under complex topological constraints. We utilize GPT-4o to classify the subject images according to the aforementioned criteria. Subsequently, all images are reviewed by five human annotators to ensure accuracy and consistency. As these images are curated based on a meticulously constructed taxonomy, they possess the comprehensiveness and representativeness required for rigorous evaluation. Given the inference efficiency of current generative models, an overabundance of redundant images would substantially inflate

Table 1: Validation of difficulty labels against model-independent image-complexity metrics. For each difficulty level, we randomly sample 50 images (5 independent runs) and report the mean ± std of Canny edge density, GLCM contrast, and JPEG bits/pixel ($Q = 85$). All three metrics increase monotonically from Easy to Hard and correlate significantly with the difficulty level.

Metric	Easy	Medium	Hard	Spearman ρ	p-value
Edge density	0.024 ± 0.004	0.048 ± 0.006	0.053 ± 0.007	0.367	3.2×10^{-6}
GLCM contrast	3.71 ± 0.20	4.07 ± 0.26	5.28 ± 0.43	0.235	2.5×10^{-3}
JPEG bits/pixel	0.71 ± 0.04	0.91 ± 0.06	0.96 ± 0.04	0.260	1.3×10^{-3}

the evaluation overhead. Ultimately, we obtain a total of **459** high-quality images.
3) Validation of Difficulty Labels against Objective Image Properties To verify that our difficulty labels reflect intrinsic image complexity rather than model-dependent heuristics, we randomly sample 50 images per difficulty level (5 independent runs) and compute three model-independent complexity metrics: Canny edge density, GLCM contrast, and JPEG bits/pixel (quality factor $Q = 85$). As reported in Tab. 1, all three metrics increase monotonically from Easy to Medium to Hard, and each exhibits a statistically significant positive Spearman correlation with the difficulty level ($p < 0.01$). This confirms that our difficulty labels correlate with objective, model-agnostic image properties, rather than relying on circular model-dependent definitions.

Prompt Generation Although DreamBench++ categorizes prompts based on their perceived difficulty, it does not provide empirical evidence to substantiate the criterion. To address this limitation, we organize the prompts according to specific application scenarios, dividing them into six categories, including (1) **Background change (BC):** scenarios involving changes in background elements. (2) **Variation in subject viewpoint or size (VS):** scenarios that entail changes in camera angle, which may include variations in subject size, lighting, or shadows. (3) **Interaction with other entities (IE):** scenarios requiring complex interactions with additional entities, potentially resulting in occlusion and necessitating adherence to physical plausibility. (4) **Attribute change (AC):** scenarios involving modifications to certain attributes of the subject, such as color or shape. (5) **Style change (SC):** scenarios involving alterations in the artistic or visual style of the subject. (6) **Imagination (IM):** scenarios where the target image depicts an imagined or fictional scene. We generate two prompts for each scenario. The specific instructions are depicted in Fig. 4. All prompts are reviewed by five human annotators to ensure they are ethical and free from defects. For the specific verification procedure, please refer to the Supplementary material (Sec. E.3). Finally, we obtain a total of **5,508** prompts. Fig. 2 shows the distribution of subject difficulty levels and prompt scenarios. We visualize the t-SNE of images from our benchmark and DreamBench++ in Fig. 3, which indicate DSH-Bench achieves superior diversity.

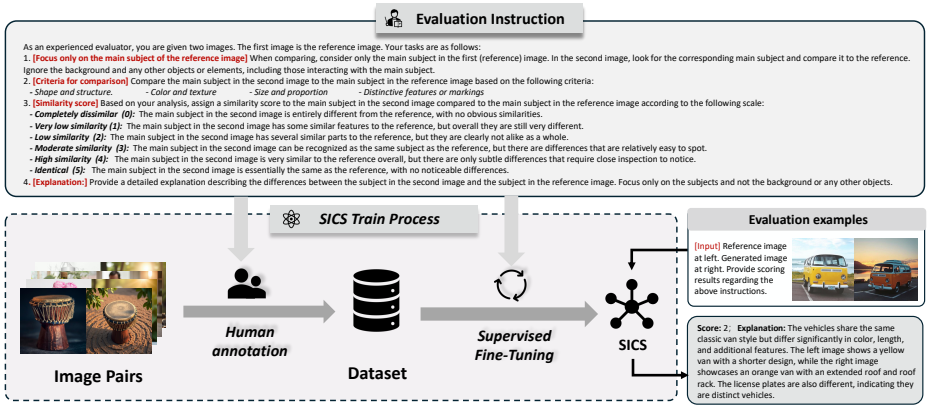


Fig. 5: The training process of SICS. We constructed and annotated a dataset specifically tailored for subject consistency determination, and subsequently trained models using this dataset.

3.2 Evaluation Dimension

Previous notable works [15, 31, 54, 63] evaluate the performance of subject-driven T2I models from two perspectives: Subject Preservation and Prompt Following. RealCustom++ [40] also uses ImageReward [73] to evaluate image quality. Therefore, DSH-Bench evaluates from the three aforementioned dimensions.

Subject Preservation DreamBench++ utilizes GPT-4o for evaluation to improve alignment with human assessments. However, the GPT-4o-based method is prohibitively expensive. To address this limitation, we propose a novel metric—**Subject Identity Consistency Score** (SICS). As shown in Fig. 5, we establish a scoring criterion for assessing subject preservation firstly. Five annotators label the collected image pairs according to the criterion. During the annotation process, each image pair is not only assigned a score but also accompanied by an explanation. Previous work [66] has indicated that labeled data with explanatory reasoning can help models better understand the underlying logic and reasoning behind the labels. We then perform meticulous fine-tuning of the model using this annotated dataset. During fine-tuning, SICS leverages prompts to explicitly prioritize subject consistency rather than global semantics, mitigating background and style artifacts that commonly bias CLIP-based approaches and yielding closer alignment with the goals of subject-consistency evaluation. Although GPT-4o demonstrates outstanding performance across a wide range of tasks, it has not been specifically optimized for subject preservation evaluation. More details of the SICS metric can be found in Supplementary material (Sec. E.2).

Prompt Following Prompt following primarily evaluates whether a model can generate images that accurately correspond to textual prompts. DreamBench++ has demonstrated that the CLIP-T score is highly consistent with human annotations. Therefore, we also adopt CLIP-T score as the evaluation metric for prompt following.

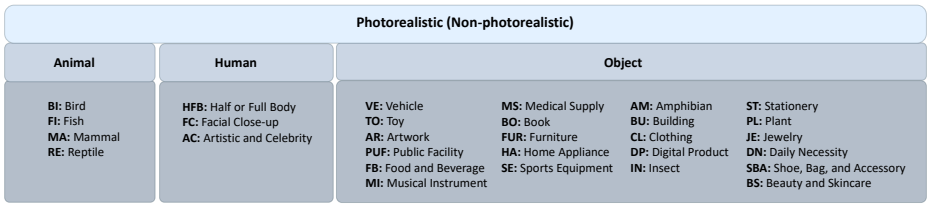


Fig. 6: Category hierarchy of the dataset. The top-level categories Photorealistic and Non-photorealistic share an identical set of sub-categories.

Image Quality HPSv2 [69] utilizes professionally annotated data to more accurately reflect human aesthetic preferences for generated images. Previous studies [59] demonstrate that models optimized with HPSv2 achieve superior performance in image quality assessment compared to existing approaches. Therefore, we adopt HPSv2 for image quality evaluation.

4 Experiment

4.1 Experiment Setup

Implementation Details We conducted experiments on the following 19 models in total: 1) Textual Inversion(TI) [16], 2) DreamBooth, 3) Custom Diffusion, 4) Hiper [19], 5) NeTI [1], 6) BLIP-Diffusion [33], 7) IP-Adapter [74], 8) MS-Diffusion [64], 9) Emu2 [58], 10) OminiControl [60], 11) SSR-Encoder [77], 12) RealCustom++ [40], 13) OmniGen [70], 14) λ-Eclipse [44], 15) UNO [68], 16) ACE++ [39], 17) DreamO [41], 18) FLUX.1 Kontext [dev] [4], 19) Nano-Banana [18]. Our experiments are conducted using the official implementations to guarantee reliability and fairness. To demonstrate that our benchmark remains challenging for advanced closed-source models, we conducted experiments using Nano-Banana. More details can be found in supplementary material (Sec. E).

Dataset Our benchmark ultimately comprises 459 subject images and 5,508 prompts. These images are distributed across 58 distinct categories. Detailed information regarding category distribution is provided in Fig. 6.

Human Annotation All annotation tasks, including labeling of the SICS training datasets, were conducted by the same five human annotators. We provide the annotators with detailed labeling guidelines and sufficient training to ensure they fully understand the subject-driven T2I generation task and could provide unbiased and discriminative scores. For additional details regarding the human annotation process, please see the supplementary material (Sec. E.4).

4.2 Main Results

SICS Results Tab. 2 presents a rigorous study of human alignment using *Kendall's τ value* (KDV) and *Spearman correlation coefficient value* (SCV)

Table 2: The human alignment degree among different metrics, measured by **Kendall's τ value** and **Spearman correlation coefficient value**. H: Human, G: GPT-4o, D: DINO, Dv2: DINOv2, CB: CLIP-B, CL: CLIP-L, S: SICS. Bold font is used to denote the maximum value in a row.

Method	Kendall↑						Spearman↑					
	H-CB	H-CL	H-D	H-Dv2	H-G	H-S	H-CB	H-CL	H-D	H-Dv2	H-G	H-S
BLIP-Diffusion	0.228	0.176	0.285	0.167	0.354	**0.531**	0.285	0.215	0.350	0.206	0.383	**0.554**
IP-Adapter	0.294	0.296	0.258	0.290	0.419	**0.622**	0.364	0.371	0.325	0.364	0.459	**0.657**
MS-Diffusion	0.158	0.090	0.116	0.122	0.119	**0.178**	**0.194**	0.109	0.144	0.156	0.131	0.189
OminiControl	0.375	0.371	0.337	0.348	0.650	**0.713**	0.490	0.486	0.441	0.453	0.729	**0.764**
SSR-Encoder	0.264	0.338	0.295	0.348	0.504	**0.664**	0.328	0.421	0.368	0.434	0.549	**0.697**
ACE++	0.341	0.325	0.312	0.312	0.425	**0.524**	0.298	0.352	0.323	0.415	0.313	**0.492**
DreamO	0.243	0.311	0.219	0.243	0.321	**0.361**	0.240	0.291	0.291	0.311	0.322	**0.383**
FLUX.1 Kontext [dev]	0.346	0.313	0.269	0.340	0.413	**0.465**	0.371	0.394	0.430	0.361	0.297	**0.496**
UNO	0.249	0.218	0.299	0.240	0.236	**0.385**	0.340	0.297	0.390	0.312	0.268	**0.426**
RealCustom++	0.181	0.128	0.206	0.241	0.291	**0.464**	0.229	0.162	0.266	0.303	0.325	**0.511**
OmniGen	0.465	0.396	0.344	0.349	0.617	**0.621**	0.579	0.497	0.440	0.456	**0.697**	0.667
λ-Eclipse	0.143	0.233	0.084	0.103	0.325	**0.375**	0.176	0.287	0.103	0.127	0.352	**0.393**
Custom Diffusion	0.316	0.336	0.382	0.425	0.487	**0.642**	0.388	0.409	0.470	0.519	0.512	**0.654**
DreamBooth	0.639	0.591	0.537	0.429	0.647	**0.692**	0.733	0.721	0.661	0.537	0.705	**0.740**
Textual Inversion	0.482	0.459	0.447	0.438	0.541	**0.568**	0.587	0.559	0.545	0.534	0.582	**0.590**
HiPer	0.338	0.387	0.351	0.404	0.584	**0.625**	0.417	0.469	0.430	0.496	0.629	**0.655**
NeTI	0.469	0.456	0.431	0.417	0.617	**0.728**	0.573	0.561	0.529	0.512	0.682	**0.778**
ALL	0.416	0.411	0.350	0.376	0.619	**0.677**	0.529	0.522	0.451	0.483	0.697	**0.734**

(metric selection rationale in supplementary material (Sec. E.2)). Notably, the image pairs used for this human-alignment study form a held-out evaluation set of 1,600 pairs that is *completely disjoint* from the 5,000 image-text pairs used to fine-tune SICS, ensuring that the reported correlations reflect genuine human alignment rather than fitting to the training data (annotation protocol detailed in supplementary material (Sec. E.4)). Our experimental results demonstrate that **SICS achieves superior alignment with human evaluations compared to existing methods**, showing consistently higher agreement across both correlation metrics in most experimental settings. Although SICS attains second-highest correlation scores in MS-Diffusion and OmniGen, it significantly outperforms GPT-4o (*GPT-4o refers to the evaluation method used in DreamBench++*) by **9.37%** (KDV) and **5.31%** (SCV). This performance gap strongly suggests SICS's enhanced capability in modeling human evaluation. Notably, GPT-4o exhibits greater consistency with human evaluation than CLIP and DINO, aligning with DreamBench++ findings. Importantly, our proposed SICS metric surpasses all existing metrics in human judgment consistency.

Quantitative & Qualitative Results Tab. 3 shows overall evaluation results. The results show that: **i) DSH-Bench poses more significant challenges than existing benchmarks.** For subject preservation and image quality, the majority of methods consistently yield lower scores on DSH-Bench. The result can be attributed to the hierarchical taxonomy sampling method employed, which allows our dataset to more accurately represent the true data distribution. Moreover, it highlights that benchmarks derived from true distributions present greater challenges. **ii)** For prompt following, DreamBench yields slightly lower scores than DSH-Bench for certain methods. In DreamBench, prompts requiring attribute change constitute 22.7%, which is higher than the 16.7% observed in DSH-Bench. Fig. 8b indicates that all methods exhibit relatively poor average

Table 3: Evaluation of Subject-driven T2I generation on open-source models. DB: DreamBench, DB++: DreamBench++, HB: DSH-Bench. All scores are normalized to 0-1. Bold indicates **the minimum value in each row** for a given evaluation dimension. Experiments show that DSH-Bench is more difficult.

Method	Subject Preservation			Prompt Following			Image Quality		
	DB	DB++	HB	DB	DB++	HB	DB	DB++	HB
BLIP-Diffusion	0.229	0.216	**0.204**	0.291	0.278	**0.277**	0.267	0.254	**0.223**
IP-Adapter	0.230	0.244	**0.229**	0.321	0.318	**0.315**	0.291	0.296	**0.266**
MS-Diffusion	**0.316**	0.346	0.352	**0.332**	0.339	0.338	0.311	0.314	**0.294**
OminiControl	0.279	0.268	**0.258**	**0.325**	0.337	0.334	0.312	0.308	**0.290**
SSR-Encoder	0.231	**0.202**	**0.202**	0.290	**0.287**	0.295	0.273	0.270	**0.247**
ACE++	0.324	0.303	**0.292**	0.321	0.316	**0.304**	0.294	0.303	**0.252**
DreamO	0.412	0.396	**0.391**	0.324	0.339	**0.326**	0.314	0.308	**0.283**
FLUX.1 Kontext [dev]	0.445	0.432	**0.424**	0.321	0.324	**0.319**	0.273	0.270	**0.288**
UNO	**0.409**	0.410	**0.409**	**0.317**	0.322	0.323	0.304	0.297	**0.278**
Emu2	0.360	0.343	**0.341**	**0.291**	0.309	0.304	0.272	0.278	**0.260**
RealCustom++	0.377	0.380	**0.375**	**0.325**	0.329	0.332	0.316	0.314	**0.298**

Table 4: DSH-Bench leaderboard. The models are ranked by the final score S_h.

Method	T2I Model	Subject Preservation	Prompt Following	Image Quality	$S_h\uparrow$
Nano-Banana	-	**0.439**	0.337	**0.302**	**0.272**
FLUX.1 Kontext [dev]	FLUX.1 Kontext	0.424	0.319	0.288	0.256
UNO	FLUX.1-dev	0.409	0.323	0.278	0.252
DreamO	FLUX.1-dev	0.391	0.326	0.283	0.251
RealCustom++	SDXL	0.375	0.332	0.294	0.251
MS-Diffusion	SDXL	0.352	**0.338**	0.294	0.248
Emu2	SDXL	0.341	0.304	0.260	0.228
OminiControl	FLUX.1-schnell	0.258	0.334	0.290	0.218
ACE++	FLUX.1-dev	0.292	0.304	0.252	0.214
IP-Adapter	SDXL	0.256	0.292	0.266	0.199
λ-Eclipse	SDXL	0.229	0.315	0.242	0.198
OmniGen	SD v1.5	0.202	0.295	0.265	0.183
SSR-Encoder	SDXL	0.188	0.322	0.247	0.181
NeTI	SD v1.4	0.192	0.301	0.234	0.176
BLIP-Diffusion	SD v1.5	0.204	0.277	0.223	0.174
DreamBooth	SD v1.5	0.158	0.321	0.245	0.164
HiPer	SD v1.4	0.135	0.318	0.247	0.151
Textual Inversion	SD v1.5	0.109	0.299	0.225	0.129
Custom Diffusion	SD v1.4	0.062	0.323	0.240	0.091

performance on prompts involving attribute change. **iii)** Tab. 4 shows that there exists a trade-off between subject preservation and prompt following. We plot the Pareto frontier (see in supplementary material (Sec. D.1)) using the data presented in Tab. 4. The primary objective is to identify a Pareto optimal solution that effectively balances the two objectives. *Additional results and discussions are in supplementary material (Sec. D.2).*

Leaderboard In order to assess a model's overall capability, we define the final score as:

$$\mathcal{S}_{\mathrm{h}} = \frac{3}{\frac{\lambda}{\mathrm{SP}} + \frac{\gamma}{\mathrm{PF}} + \frac{\mu}{\mathrm{IQ}}} \tag{1}$$

SP, PF, and IQ represent the scores for Subject Preservation, Prompt Following, and Image Quality, respectively. λ, γ, μ are the weights assigned to the importance of each corresponding dimension. In this study, we set $\lambda = 1.5, \gamma = 1.5, \mu = 1$,

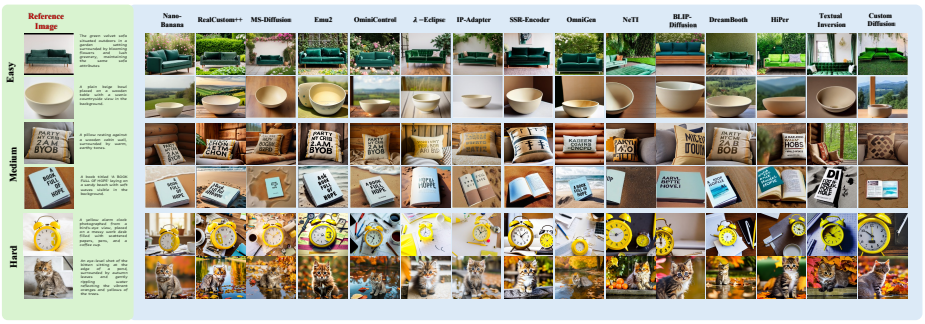


Fig. 7: Examples generated by methods listed in the leaderboard.

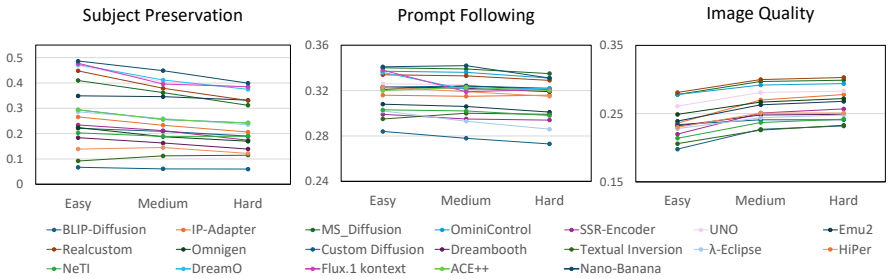
as subject preservation and prompt following are of paramount importance in subject-driven T2I generation. The harmonic mean requires strong performance across all dimensions to yield a high overall score. We rank models by S_h scores in Tab. 4. Nano-Banana exhibits relatively strong overall performance. Among open-source models, FLUX.1 Kontext [dev] performs best.

5 Analysis

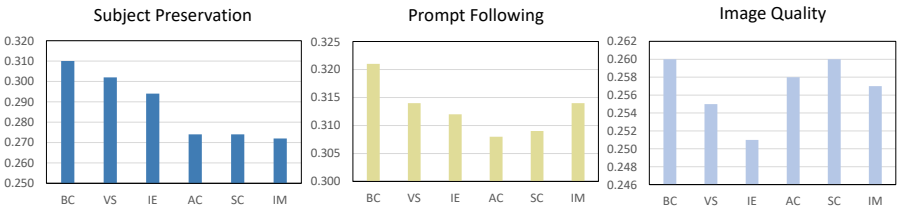
In this section, we conduct a detailed analysis of the performance of all methods based on the hierarchical category, subject difficulty level, and prompt scenario. Fig. 7 shows qualitative examples generated by the methods in the leaderboard.

A scientific and comprehensive subject image sampling method is necessary Fig. 8c present the performance of various methods in the third-level categories. The results reveal that model robustness varies considerably among categories. For example, performance in category "*Book*" (both photorealistic and non-photorealistic) is substantially lower. This disparity suggests that the absence of subject images from specific categories can lead to biased evaluation results, highlighting the importance of data diversity. Furthermore, Fig. 8c also demonstrates that none of the current models perform well across all categories. We hypothesize that this may be related to the varying complexity of the subjects within different categories. A more detailed analysis of model performance in different categories can be found in supplementary material (Sec D.1).

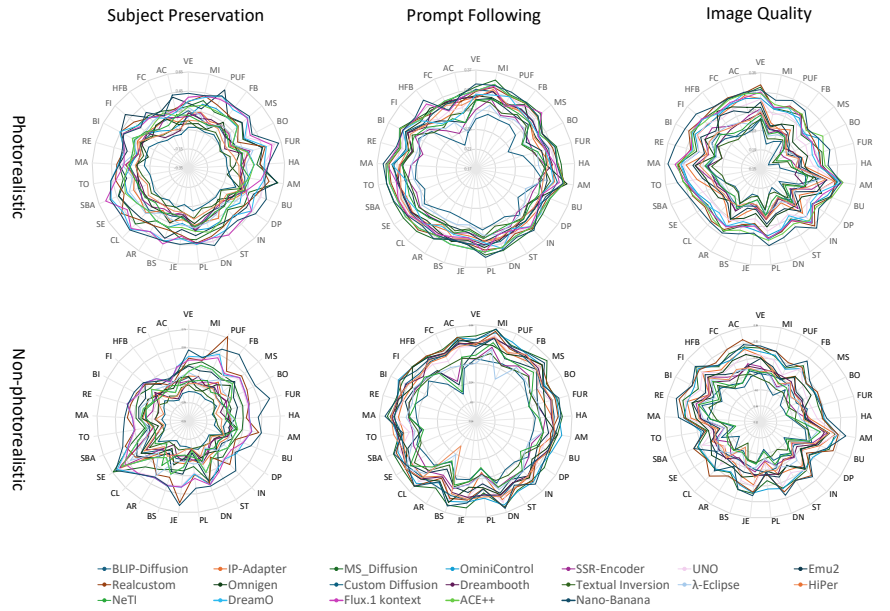
Current subject-driven T2I models exhibit performance degradation on hard level subjects As illustrated in Fig. 8a, the model exhibits substantial variation in performance across different difficulty levels: 1) For subject preservation, there is a pronounced decline in performance as the difficulty of the subject images increases. The model achieves significantly better results on images classified as simple compared to those categorized as hard. This observation supports the validity of our image difficulty classification scheme. 2) As demonstrated in Tab. 4, current closed-source models have achieved remarkable advancements in subject-driven T2I tasks, yielding highly competitive scores in our benchmark evaluation. Nevertheless, as illustrated in Fig. 8c, the performance



(a) DSH-Bench scores across different subject difficulty level.



(b) DSH-Bench scores across different prompt scenarios.



(c) DSH-Bench scores across different categories.

Fig. 8: Comparison for DSH-Bench scores across different evaluation dimensions. Note that the categories are divided into two types: Photorealistic and Non-photorealistic. The specific metric values are provided in the supplementary material (Sec. D.2). Category definitions refer to supplementary material (Sec. E.3). Best viewed when zoomed in.

of Nano-Banana on challenging categories leaves ample room for improvement. Thus, our benchmark retains its formidable challenge. 3) For prompt following, Fig. 8a shows that model capability is minimally influenced by the subject difficulty level. This could be explained by the fact that CLIP-T primarily emphasizes overall semantic information. Hence, as long as the generated image correctly represents the general category and overall shape, the evaluation score is unlikely to be substantially reduced, even if finer details are not perfectly captured. *Given these findings, it is crucial to enhance models' ability to encode and reconstruct complex subject details more effectively in future research.*

The subject-driven T2I capability for different prompt scenarios is not robust Fig. 8b shows the average performance of all models across six prompt scenarios. The results show that: 1) In BC, VS, and IE scenarios, the model's performance consistently declines across all evaluation dimensions. This trend suggests that scenario difficulty increases progressively from BC to IE. Notably, the finding that the IE scenario is more challenging than the BC scenario aligns with intuitive expectations. 2) For subject preservation, the model's average performance across the AC, SC, and IM scenarios remains relatively low. This could be because the generated subjects undergo partial modifications relative to the original subjects in these three scenarios. *Given these findings, more emphasis should be placed on enhancing methods for IE prompt scenario. For instance, increasing the volume of training data tailored to these specific contexts.*

6 Conclusion

This paper introduces a novel benchmark called DSH-Bench, designed specifically for subject-driven T2I generation. Key features include: 1) a hierarchical category system in image collection to ensure both the diversity and comprehensiveness of subject images; 2) an innovative classification scheme for categorizing subject difficulty levels and prompt scenarios to obtain valuable insights; and 3) a human-aligned and more efficient metric for subject preservation. The benchmark is publicly available at `https://github.com/sunriverhzy/DSH-Bench` to support the advancement in future research.

References

1. Alaluf, Y., Richardson, E., Metzer, G., Cohen-Or, D.: A neural space-time representation for text-to-image personalization. ACM Transactions on Graphics (TOG) **42**(6), 1–10 (2023)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), `https://arxiv.org/abs/2502.13923`
3. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S., et al.: ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)

4. Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv e-prints pp. arXiv–2506 (2025)
5. 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)
6. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K.P., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. In: International Conference on Machine Learning. pp. 4055–4075. PMLR (2023)
7. Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. Advances in Neural Information Processing Systems **36**, 30286–30305 (2023)
8. Chen, Z., Fang, S., Liu, W., He, Q., Huang, M., Zhang, Y., Mao, Z.: Dreamidentity: Improved editability for efficient face-identity preserved image generation (2023), https://arxiv.org/abs/2307.00300
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. Advances in neural information processing systems **34**, 19822–19835 (2021)
11. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al.: Dreamllm: Synergistic multimodal comprehension and creation. In: ICLR (2024)
12. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc. (2016), https://proceedings.neurips.cc/paper_files/paper/2016/file/371bce7dc83817b7893bcdeed13799b5-Paper.pdf
13. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In: Advances in Neural Information Processing Systems. vol. 36, pp. 50742–50768 (2023)
14. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: European Conference on Computer Vision. pp. 89–106. Springer (2022)
15. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion (2022). https://doi.org/10.48550/ARXIV.2208.01618, https://arxiv.org/abs/2208.01618
16. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=NAQvF08TcyG
17. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Encoder-based domain tuning for fast personalization of text-to-image models. ACM Transactions on Graphics (TOG) **42**(4), 1–13 (2023)

18. Google: Introducing gemini 2.5 flash image, our state-of-the-art image model (2025), https://developers.googleblog.com/introducing-gemini-2-5-flash-image/, accessed: 2025-12-15
19. Han, I., Yang, S., Kwon, T., Ye, J.C.: Highly personalized text embedding for image manipulation by stable diffusion. arXiv preprint arXiv:2303.08767 (2023)
20. Hao, S., Han, K., Zhao, S., Wong, K.Y.K.: Vico: Plug-and-play visual condition for personalized text-to-image generation. arXiv preprint arXiv:2306.00971 (2023)
21. He, J., Tuo, Y., Chen, B., Zhong, C., Geng, Y., Bo, L.: Anystory: Towards unified single and multiple subject personalization in text-to-image generation (2025), https://arxiv.org/abs/2501.09503
22. Hu, H., Chan, K.C.K., Su, Y.C., Chen, W., Li, Y., Sohn, K., Zhao, Y., Ben, X., Gong, B., Cohen, W., Chang, M.W., Jia, X.: Instruct-imagen: Image generation with multi-modal instruction (2024), https://arxiv.org/abs/2401.01952
23. Hu, H., Chan, K.C., Su, Y.C., Chen, W., Li, Y., Sohn, K., Zhao, Y., Ben, X., Gong, B., Cohen, W., et al.: Instruct-imagen: Image generation with multi-modal instruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4754–4763 (2024)
24. Hua, H., Zeng, Z., Song, Y., Tang, Y., He, L., Aliaga, D., Xiong, W., Luo, J.: Mmig-bench: Towards comprehensive and explainable evaluation of multi-modal image generation models. Advances in Neural Information Processing Systems **38** (2025)
25. Hua, M., Liu, J., Ding, F., Liu, W., Wu, J., He, Q.: Dreamtuner: Single image is enough for subject-driven generation (2023), https://arxiv.org/abs/2312.13691
26. Huang, L., Lin, H., Zhou, Y., Xiao, K.: Flexip: Dynamic control of preservation and personality for customized image generation (2025), https://arxiv.org/abs/2504.07405
27. Huang, Z., Zhuang, S., Fu, C., Yang, B., Zhang, Y., Sun, C., Zhang, Z., Wang, Y., Li, C., Zha, Z.J.: Wegen: A unified model for interactive multimodal generation as we chat (2025), https://arxiv.org/abs/2503.01115
28. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10124–10134 (2023)
29. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
30. 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)
31. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1931–1941 (2023)
32. Le, D.H., Pham, T., Lee, S., Clark, C., Kembhavi, A., Mandt, S., Krishna, R., Lu, J.: One diffusion to generate them all (2024), https://arxiv.org/abs/2411.16318
33. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. Advances in Neural Information Processing Systems **36**, 30146–30166 (2023)
34. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked id embedding (2023), https://arxiv.org/abs/2312.04461

35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. pp. 740–755. Springer (2014)
36. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2125–2134 (2021)
37. Liu, Z., Zhang, Y., Shen, Y., Zheng, K., Zhu, K., Feng, R., Liu, Y., Zhao, D., Zhou, J., Cao, Y.: Cones 2: Customizable image synthesis with multiple subjects (2023), https://arxiv.org/abs/2305.19327
38. Ma, J., Liang, J., Chen, C., Lu, H.: Subject-diffusion:open domain personalized text-to-image generation without test-time fine-tuning (2024), https://arxiv.org/abs/2307.11410
39. Mao, C., Zhang, J., Pan, Y., Jiang, Z., Han, Z., Liu, Y., Zhou, J.: Ace++: Instruction-based image creation and editing via context-aware content filling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1958–1966 (2025)
40. Mao, Z., Huang, M., Ding, F., Liu, M., He, Q., Zhang, Y.: Realcustom++: Representing images as real-word for real-time customization. arXiv e-prints pp. arXiv–2408 (2024)
41. Mou, C., Wu, Y., Wu, W., Guo, Z., Zhang, P., Cheng, Y., Luo, Y., Ding, F., Zhang, S., Li, X., et al.: Dreamo: A unified framework for image customization. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–12 (2025)
42. OpenAI: Introducing gpt-4o and more tools to chatgpt free users. https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/ (2024), accessed: 2024-06-15
43. Patashnik, O., Gal, R., Ostashev, D., Tulyakov, S., Aberman, K., Cohen-Or, D.: Nested attention: Semantic-aware attention values for concept personalization (2025), https://arxiv.org/abs/2501.01407
44. Patel, M., Jung, S., Baral, C., Yang, Y.: λ-eclipse: Multi-concept personalized text-to-image diffusion models by leveraging clip latent space. arXiv preprint arXiv:2402.05195 (2024)
45. Peebles, W., Xie, S.: Scalable diffusion models with transformers (2023), https://arxiv.org/abs/2212.09748
46. Peng, Y., Cui, Y., Tang, H., Qi, Z., Dong, R., Bai, J., Han, C., Ge, Z., Zhang, X., Xia, S.T.: Dreambench++: A human-aligned benchmark for personalized image generation. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=4GSOESJrk6
47. 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 (2023), https://arxiv.org/abs/2307.01952
48. Prashnani, E., Cai, H., Mostofi, Y., Sen, P.: Pieapp: Perceptual image-error assessment through pairwise preference. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1808–1817 (2018). https://doi.org/10.1109/CVPR.2018.00194
49. Qiu, Z., Liu, W., Feng, H., Xue, Y., Feng, Y., Liu, Z., Zhang, D., Weller, A., Schölkopf, B.: Controlling text-to-image diffusion by orthogonal finetuning. Advances in Neural Information Processing Systems **36**, 79320–79362 (2023)
50. 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)
51. 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)
52. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022), https://arxiv.org/abs/2112.10752
53. Rowles, C., Vainer, S., Nigris, D.D., Elizarov, S., Kutsy, K., Donné, S.: Ipadapter-instruct: Resolving ambiguity in image-based conditioning using instruct prompts (2024), https://arxiv.org/abs/2408.03209
54. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
55. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
56. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning (2023), https://arxiv.org/abs/2304.03411
57. Slobodkin, A., Taitelbaum, H., Bitton, Y., Gordon, B., Sokolik, M., Guetta, N.B., Gueta, A., Rassin, R., Lischinski, D., Szpektor, I.: Refvnli: Towards scalable evaluation of subject-driven text-to-image generation. arXiv preprint arXiv:2504.17502 (2025)
58. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14398–14409 (2024)
59. Sun, S., Qu, B., Liang, X., Fan, S., Gao, W.: Ie-bench: Advancing the measurement of text-driven image editing for human perception alignment. arXiv preprint arXiv:2501.09927 (2025)
60. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. arXiv preprint arXiv:2411.15098 (2024)
61. Unsplash: Unsplash. https://unsplash.com/, accessed: 2026-06-23
62. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: P+: Extended textual conditioning in text-to-image generation (2023), https://arxiv.org/abs/2303.09522
63. Wang, H., Spinelli, M., Wang, Q., Bai, X., Qin, Z., Chen, A.: Instantstyle: Free lunch towards style-preserving in text-to-image generation. arXiv preprint arXiv:2404.02733 (2024)
64. Wang, X., Fu, S., Huang, Q., He, W., Jiang, H.: Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. arXiv preprint arXiv:2406.07209 (2024)
65. Wang, Y., Zang, Y., Li, H., Jin, C., Wang, J.: Unified reward model for multimodal understanding and generation. arXiv preprint arXiv:2503.05236 (2025)
66. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)

67. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15943–15953 (2023)
68. Wu, S., Huang, M., Wu, W., Cheng, Y., Ding, F., He, Q.: Less-to-more generalization: Unlocking more controllability by in-context generation. arXiv preprint arXiv:2504.02160 (2025)
69. Wu, X., Sun, K., Zhu, F., Zhao, R., Li, H.: Human preference score: Better aligning text-to-image models with human preference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2096–2105 (2023)
70. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. arXiv preprint arXiv:2409.11340 (2024)
71. Xiong, Z., Xiong, W., Shi, J., Zhang, H., Song, Y., Jacobs, N.: Groundingbooth: Grounding text-to-image customization (2025), https://arxiv.org/abs/2409.08520
72. 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. arXiv preprint arXiv:2412.21059 (2024)
73. 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)
74. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
75. Zeng, Y., Patel, V.M., Wang, H., Huang, X., Wang, T.C., Liu, M.Y., Balaji, Y.: Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation (2024), https://arxiv.org/abs/2407.06187
76. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068
77. Zhang, Y., Song, Y., Liu, J., Wang, R., Yu, J., Tang, H., Li, H., Tang, X., Hu, Y., Pan, H., et al.: Ssr-encoder: Encoding selective subject representation for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8069–8078 (2024)