DREAMBENCH++: A HUMAN-ALIGNED BENCHMARK FOR PERSONALIZED IMAGE GENERATION

Anonymous authors
Paper under double-blind review

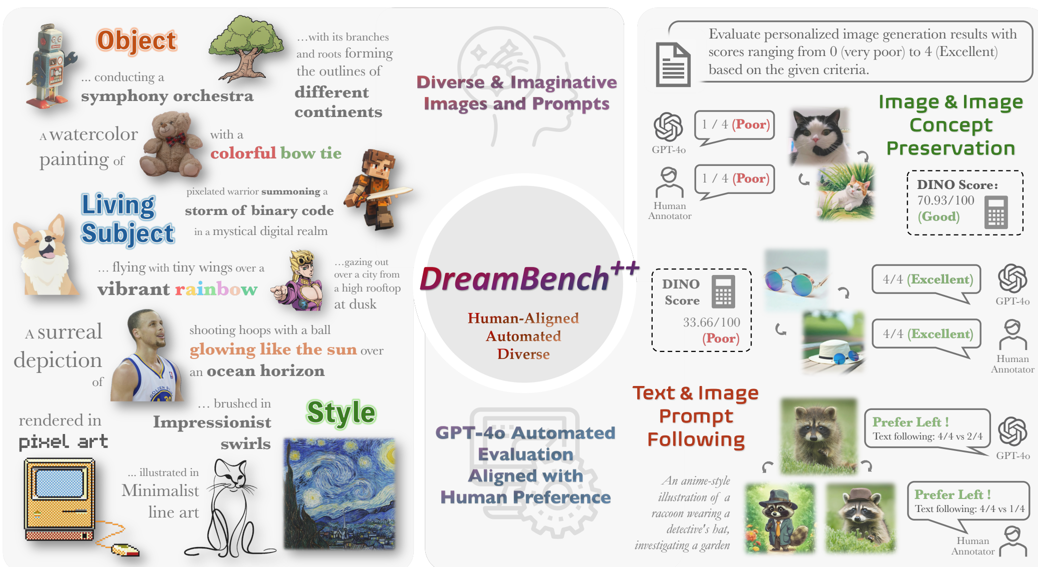


Figure 1: **Overview of DREAMBENCH++.** We collect diverse images and prompts, and utilize GPT-4o for automated evaluation aligned with human preference.

ABSTRACT

Personalized image generation holds great promise in assisting humans in everyday work and life due to its impressive function in creatively generating personalized content. However, current evaluations either are automated but misalign with humans or require human evaluations that are time-consuming and expensive. In this work, we present DREAMBENCH++, a human-aligned benchmark that advanced multimodal GPT models automate. Specifically, we systematically design the prompts to let GPT be both human-aligned and self-aligned, empowered with task reinforcement. Further, we construct a comprehensive dataset comprising diverse images and prompts. By benchmarking 7 modern generative models, we demonstrate that DREAMBENCH++ results in significantly more human-aligned evaluation, helping boost the community with innovative findings.

1 INTRODUCTION

Driven by the significant advances in large-scale text-to-image (T2I) generative models (Rombach et al., 2022; Ramesh et al., 2021; Betker et al., 2023; Ramesh et al., 2022; Nichol et al., 2022; Saharia et al., 2022b; Yu et al., 2022; Chang et al., 2023; Gafni et al., 2022; Ding et al., 2021; 2022; Balaji et al., 2022; Kang et al., 2023; Dong et al., 2024), it is now possible to generate images conditioned on not only arbitrary text prompts but also by given reference images—*personalized* image generation (Ruiz et al., 2023; Gal et al., 2023a; Li et al., 2023a; Ye et al., 2023; Kumari et al., 2023; Gal et al., 2023b; Arar et al., 2023; Chen et al., 2023c; Jia et al., 2023; Chen et al., 2023a; Xiao et al., 2023; Tewel et al., 2023; Wei et al., 2023; Ma et al., 2023; Hua et al., 2023; Wang et al., 2024b; Lv et al., 2024; Wang et al., 2024a; Chen et al., 2023b; Tumanyan et al., 2023; Zhou et al.,

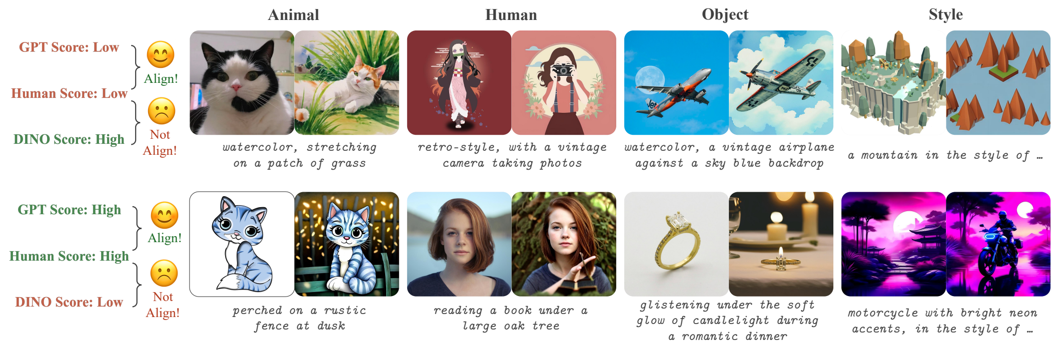


Figure 2: **Qualitative comparison of concept preservation evaluation** between DREAMBENCH++ and traditional DINO Caron et al. (2021). DINO often fails to yield human-aligned evaluation while our DREAMBENCH++ succeeds using multimodal GPT models as the evaluator.

2024; He et al., 2024c; Wang et al., 2024c; Wu et al., 2024a; He et al., 2024a; Xiao et al., 2024; Arar et al., 2024; Huang et al., 2024b;a; Pang et al., 2024a;b; Qiu et al., 2023; Hu et al., 2024). In general, to be useful as an artistic creation tool for inspiration or products (Yacoubian, 2022), the following two basic criteria must be fulfilled: **i) Prompt following** (image & prompt consistency). Generated images must follow the prompt description, which is a requirement shared with vanilla T2I generation (Betker et al., 2023; Ramesh et al., 2022). **ii) Concept preservation** (image & image consistency). For personalized image generation, the concept of the reference image, *i.e.*, the main subject's semantic details (*e.g.*, facial characters) or high-level abstractions (*e.g.*, overall style), must be preserved. For example, a user may want to "imagine his own dog traveling around the world" (Ruiz et al., 2023), and the generated dog must be the same as his but traveling.

To meet the aforementioned requirements, numerous efforts have been devoted. One line of fine-tuning-based works focuses on fine-tuning general T2I models to specialist personalization models by reproducing specific concepts present in training sets (Ruiz et al., 2023; Gal et al., 2023a; Chen et al., 2023c; Kumari et al., 2023). Meanwhile, another line of encoder-based works, instead, achieves concept-preservation by training features adaptation to inject reference image features into a general T2I model (Ye et al., 2023; Gal et al., 2023b; Arar et al., 2023; Dong et al., 2024; Sun et al., 2024a;b; Pan et al., 2024). Despite remarkable progress, one question arises: *can we comprehensively evaluate these models to figure out which technical route is superior and where to head?*

In this work, we aim to answer this question by developing a new benchmark that properly evaluates personalized T2I models driven by the above two requirements. We present DREAMBENCH++, a comprehensive benchmark designed based on the following *de-facto* principled advantages:

1. **Human-Aligned** As shown in Fig. 2, traditional metrics like DINO (Caron et al., 2021) and CLIP (Radford et al., 2021) often result in significant discrepancies from humans. This is caused by the image similarity measurement nature of DINO and CLIP models, and thus crowd-sourced *human evaluation* is typically necessary for obtaining a correct *quantitative* understanding of generated images (Lee et al., 2023; Ku et al., 2024; Xu et al., 2023). Therefore, different from existing works that utilize CLIP and DINO as metrics that may be humanly misaligned, our DREAMBENCH++ demonstrates surprisingly consistent evaluation results aligned with humans. For instance, by evaluating 7 modern models, DREAMBENCH++ achieves **79.64%** and **93.18%** agreement with human's evaluation in concept preservation and prompt following capabilities, respectively. Notably, it is **+54.1%** and **+50.7%** higher than traditional DINO and CLIP metrics.
2. **Automated** However, it is non-standardized and expensive to perform high-quality human evaluations. To address this challenge, DREAMBENCH++ achieves automated but human-aligned evaluation by using advanced multimodal GPT models such as GPT-4o (OpenAI, 2024) as metrics. The challenges lie in two aspects: i) prompt design and ii) reasoning procedure for scoring. We systematically standardize the automated GPT evaluation by first designing the *evaluation instructions* that provide overall task requirements, where language is a general interface for instructing human preference. Inspired by Self-Align (Sun et al., 2023), we instruct GPT to conduct *internal thinking* that aligns itself for better task and preference understanding. Then, GPT provides the *summary & planning* for the task and scoring criteria, and the final scores are provided with optional *chain-of-thought (CoT)* (Wei et al., 2022; Zhang et al., 2023d).

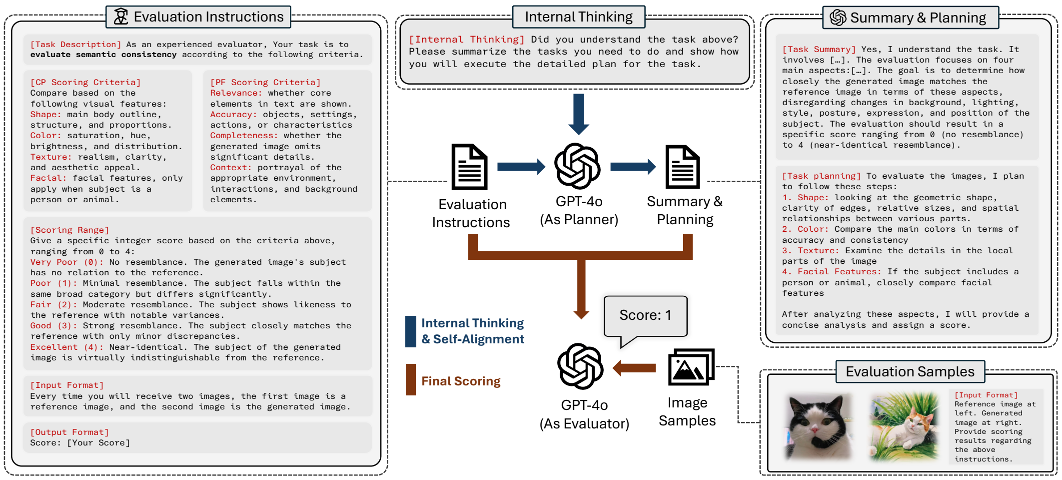


Figure 3: **Overall procedure of prompting GPT-4o for automated evaluation.** The evaluation instructions are meta-prompting information written by humans, including task description, scoring criteria, scoring range, and format specification. Then, GPT-4o is prompted with reasoning instructions to perform internal thinking that provides a self-aligned task summary and planning. Finally, all prompts and reasoning outputs are joined with image samples for score outputs.

3. **Diverse** To avoid bias from low-diversity evaluation data, DREAMBENCH++ compiles a wide range of images, covering varying levels of difficulty from simpler animals and styles to more complex human subjects, objects, and non-natural styles (see Fig. 1). Unlike DreamBench (Ruiz et al., 2023), which includes only 30 subjects and 25 prompts, DREAMBENCH++ significantly expands the dataset to 150 images and 1,350 prompts—**5**× and **54**× more, respectively. While CustomConcept101 (Kumari et al., 2023) offers 101 subjects, its diversity is limited by repetitive image categories and a focus on photorealistic styles, with simple prompts that restrict its ability to evaluate models on more complex tasks. Consequently, DREAMBENCH++ enables more robust and comprehensive conclusions in model evaluation.

Takeaways We present some insightful findings from evaluating seven modern personalized T2I models: i) DINO-based ratings prioritize overall shape and color over detailed features, making them suboptimal for evaluating personalized image generation; ii) The primary goal is to achieve a Pareto optimal balance between concept preservation and prompt adherence. Among the models, DreamBooth (Ruiz et al., 2023) excels in preserving detailed visual features while closely following text prompts; iii) Current models perform well in animal and style categories but struggle with human images due to sensitivity to facial details and diverse object categories. While existing work (Wang et al., 2024b; Valevski et al., 2023; Yan et al., 2023; Ye et al., 2023; Xiao et al., 2023) addresses facial feature preservation, the challenge of object diversity remains underexplored.

We are presenting DREAMBENCH++ with open-sourced codes and evaluation standardization to promote innovation within the research community. In addition, we believe our design of the human-aligned & automated evaluation using advanced foundation models is robust and transferrable to other domains and foundation models (*e.g.*, GPT-5 in the future).

2 DREAMBENCH++

We introduce DREAMBENCH++, a human-aligned, automated, and diverse benchmark that evaluates the two capabilities of personalized image generation models. In the following, we describe how we construct DREAMBENCH++ from two aspects: prompts and data.

2.1 PROMPTING GPT FOR AUTOMATED & HUMAN-ALIGNED BENCHMARKING

It is challenging to obtain a solid quantitative understanding of generated models, especially when evaluating visual contents that rely on human evaluations (Lee et al., 2023; Ku et al., 2024). Thus, it is critical to achieve automated evaluation by utilizing multimodal GPT models, which are trained

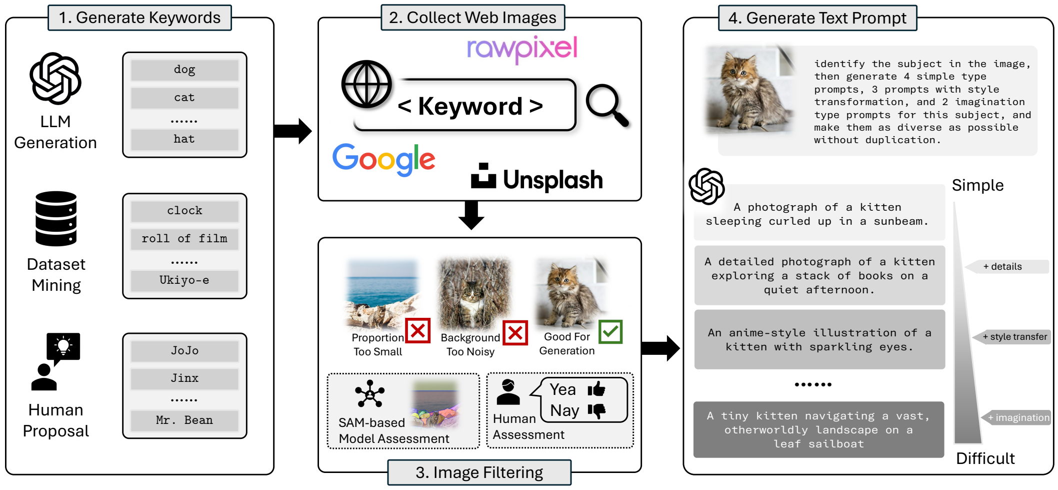


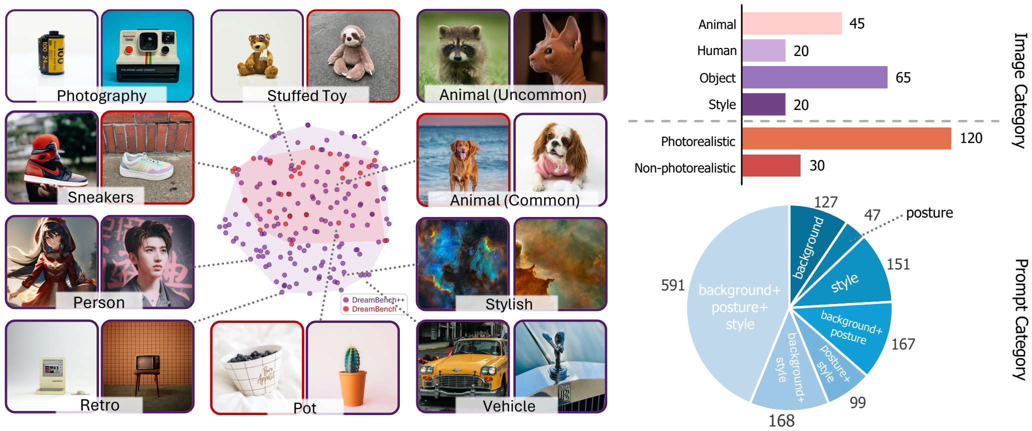
Figure 4: **Dataset construction process of our DreamBench++.** We start by obtaining keywords through GPT generation, existing datasets, and human proposals. Next, we collect corresponding images from the internet. These images are then filtered to remove low-quality ones through both model and human assessment. The remaining high-quality images are used as input for GPT-4o to generate text prompts of varying difficulty levels.

particularly in the principle of aligning with human preference (Ouyang et al., 2022; Christiano et al., 2017; OpenAI, 2024; 2023). This is evidenced by the recent progress achieved by Wu et al., which demonstrates that GPT-4V (OpenAI, 2023) can serve as a human-aligned text-to-3D generation evaluator. However, as pointed out by Zhang et al. and Ku et al., multimodal GPT models often fall short in evaluating personalized image generation—often more challenging when distinguishing *subtle difference* for concept preservation assessment using GPT—still underexplored. To tackle this issue, we detail how we systematically design the prompt of multimodal GPT (GPT-4o (OpenAI, 2024), by default) for human alignment reinforcement but also improve the reasoning progress that helps the GPT models to be more self-aligned, introduced as follows.

Compare or rate? There are typically two schemes for quantitatively evaluating generative models in human evaluations: *rating* and *comparison* (Zhang et al., 2023c; Zheng et al., 2023). The rating scheme requires human reviewers to assign an absolute score to each instance, while the comparison scheme asks human reviewers to express a relative preference among different instances. Though effective as the comparison scheme is when humans are involved, we find that there are two critical issues. **i)** ***Positional Bias***: the scoring results of GPT-4V/GPT-4o is sensitive to the order in which images are presented (OpenAI, 2023; Wang et al., 2023a;b; Zhang et al., 2023c; Wu et al., 2024b; Zheng et al., 2023), making it unsuitable for comparison scheme. **ii)** ***Quadratic Complexity***: As the number of methods increases, the number of essential evaluation runs for numerical rating increases linearly, while the number of comparative assessments increases quadratically. Therefore, direct numerical rating is more efficient and scalable when evaluating multiple methods. Hence, in this work, we adhere to the rating scheme, and we establish a *5-level rating scheme* where scores are integers ranging from 0 (very poor) to 4 (excellent).

Evaluation Instructions The evaluation instructions serve as the meta-prompting that describes overall tasks, which is shown in Fig. 3. As stated in Section 1, there are two fundamental quality criteria to be evaluated: i) *concept preservation* and ii) *prompt following*. For each aspect, we use a similar prompt template that contains ❶ **task description**, ❷ **scoring criteria explanation**, ❸ **scoring range definition**, and ❹ **format specification**. Only the scoring criteria are tailored for different tasks: for concept preservation evaluation, we prompt GPT to focus on *shape*, *color*, *texture*, and *facial features* (if applicable), while for prompt following evaluation we requested for focus on *relevance*, *accuracy*, *completeness* and *context*.

Reasoning Instructions Given the evaluation instructions, it is crucial to reinforce the alignment with both the human instruction and itself to largely leverage the pretrained knowledge. To this end, we adopt a 2-step evaluation policy as follows: **i)** ***Internal Thinking***: Inspired by Self-Align (Sun et al., 2023), we introduce internal thinking to strengthen task understanding and instruction follow-



(a) t-SNE Visualized Data Distribution Comparison of DreamBench and DreamBench++

(b) Image & Prompt Category Distribution of DreamBench++

Figure 5: **Data distribution visualization.** (a) Images comparison between DreamBench and DreamBench++ using t-SNE (Van der Maaten & Hinton, 2008; Poličar et al., 2019). (b) Image and prompt distribution of DreamBench++.

ing capabilities. Specifically, we prompt the GPT model by asking if it understands the task or not and let it summarize the task. **ii)** ***Summary & Planning***: According to the given internal thinking instruction, the GPT will summarize and plan for the evaluation task itself. It can also be viewed as a generalized form of chain-of-thought reasoning (Wei et al., 2022; Zhang et al., 2023d). The complete procedure is illustrated in Fig. 3.

2.2 Scaling Up Personalized Image Generation Benchmarking

Pioneering works like DreamBooth (Ruiz et al., 2023), SuTI (Chen et al., 2023c) and CustomConcept101 (Kumari et al., 2023) have successfully set up baseline datasets for the evaluation of personalized image generation, and DreamBench++ follows them to categorize images into three types: ❶ **objects**, ❷ **living subjects**, and ❸ **styles**. However, due to the small-scale nature of DreamBench and the limited diversity of CustomConcept101, it is limited as some methods may converge well on its samples while performing unsatisfactorily on other data. To avoid this possible biased evaluation, we scale up the benchmarking data by increasing both image numbers and diversity.

Data Construction from Internet There are broad images on the Internet, and many datasets are constructed from it (Schuhmann et al., 2021; Jia et al., 2021). DreamBench++ mainly collects images from Unsplash (uns), Rawpixel (raw), and Google Image Search (goo), along with contributions from individuals with authorized permissions. *Each image's copyright status has been verified for academic suitability.* As shown in Fig. 4, we collect and construct high-quality data in DreamBench++ by following 3 steps:

- **Keywords Generation** First, we generate 200 relevant keywords using GPT-4o and join them with the 200 most frequent keywords from Unsplash. After filtering out duplicated keywords, seven human annotators will extend the list to around 300 based on their interests.
- **Internet Images Collection** Given selected keywords, we retrieved corresponding images from Unsplash, Rawpixel, and Google Image Search. To filter out images unsuitable for personalized image generation, SAM (Kirillov et al., 2023) is applied to identify subject regions in images and discard those with too small subject areas. Human annotators will then filter out images with noisy backgrounds. Curated images were cropped to centralize the subject, resulting in two images per keyword. Keywords that fail to yield suitable images will be discarded in this process.
- **Prompt Generation** After image collection, 9 text prompts per image were generated using GPT-4o, designed to cover a range of difficulties: 4 prompts for ❶ **photorealistic** styles, 3 for ❷ **non-photorealistic** styles, and 2 for ❸ **complicated & imaginative** contents. To align with established evaluation methods, we use few-shot prompts selected from PartiPrompts (Yu et al., 2022). Human calibration ensures that all generated prompts are ethical and without flaws. As a result, the construction process finally yields 150 high-quality images and 1,350 prompts.

Table 1: **Evaluation of personalized image generation models on DreamBench++.** All scores are normalized to 0-1, and -I & -T represent image & text, respectively.

Method	T2I Model	Concept Preservation				Prompt Following		
		Human	GPT	DINO-I	CLIP-I	Human	GPT	CLIP-T
• Textual Inversion	SD v1.5	0.316	0.378±0.0012	0.437	0.726	0.604	0.624±0.0033	0.302
• DreamBooth	SD v1.5	0.453	0.493±0.0012	0.544	0.753	0.679	0.721±0.0016	0.323
• DreamBooth LoRA	SDXL v1.0	0.571	0.597±0.0007	0.628	0.784	0.821	0.865±0.0007	0.341
• BLIP-Diffusion	SD v1.5	0.513	0.547±0.0010	0.649	0.823	0.577	0.495±0.0005	0.286
• Emu2	SDXL v1.0	0.410	0.528±0.0016	0.539	0.763	0.641	0.689±0.0010	0.310
• IP-Adapter-Plus ViT-H	SDXL v1.0	0.755	0.833±0.0008	0.834	0.917	0.541	0.413±0.0005	0.282
• IP-Adapter ViT-G	SDXL v1.0	0.570	0.593±0.0018	0.667	0.855	0.688	0.640±0.0017	0.309

Diversity Visualization Internet images are numerous. However, there is a bias towards *photorealistic* styles. To diversify, various *non-photorealistic* styles are enlisted, and human annotators are tasked to gather images for each style, including *anime*, *sketches*, *traditional Chinese paintings*, *artworks*, and *cartoon characters from games*. Then, a manual selection process ensures a balanced distribution across subject classes and between photorealistic and non-photorealistic styles. In Fig. 5(a), we visualize the t-SNE (Van der Maaten & Hinton, 2008; Poličar et al., 2019) of images from DreamBench and DreamBench++, which demonstrates the superiority of DreamBench++ in diversity. Besides, Fig. 5(b) presents the detailed image distribution in DreamBench++.

3 Experiments

3.1 Experimental Setup

Reimplementation Details We conduct experiments on two mainstream methods: i) • ***Fine-tuning-based methods***, including ❶ **Textual Inversion (TI) (Gal et al., 2023a)**, ❷ **DreamBooth (Ruiz et al., 2023)**, and ❸ **DreamBooth LoRA (DreamBooth-L) (Ruiz et al., 2023; Hu et al., 2022)**; ii) • ***Encoder-based methods*** that trains feature adaptation, including ❹ **BLIP-Diffusion (BLIP-D) (Li et al., 2023a)**, ❺ **Emu2 (Sun et al., 2024a)**, ❻ **IP-Adapter-Plus ViT-H (IP-Adapt.-P) (Ye et al., 2023)**, and ❼ **IP-Adapter ViT-G (IP-Adapt.) (Ye et al., 2023)**. All methods are based on base T2I models, including SD v1.5 (Rombach et al., 2022) and SDXL v1.0 (Podell et al., 2024). We stay true to the official implementations wherever possible and dedicate significant effort to parameter tuning for performance assurance on DreamBench, see Appendix B.

Human Annotators We employ 7 human annotators to score each instance in DreamBench++ to obtain ground truth human preference data. We provide human annotators with sufficient training to ensure they fully understand the personalized T2I generation task and can provide *unbiased* and *discriminating* scores. *The scoring task and scheme given to humans are identical to those used for GPT, as described in Section 2.* The GPT results and human results are isolated to avoid hindsight bias. Additionally, we ensure that each instance is rated by *at least two humans* to reduce noise.

Figure 6: **Comparison between images of high DINO score and high GPT-4o score.** Instances with high human scores are ticked, and those with low human scores are crossed. DINO tends to yield high scores to images that preserve overall shape but do not put much weight on color, texture, and facial features, leading to frequent contradiction with human preference.

Table 2: **DreamBench++ leaderboard.** Both scores for concept preservation (CP) and prompt following (PF) are presented and divided by established concept and prompt categories. The models are ranked by the product of CP and PF scores (CP·PF).

Method	T2I Model	Concept Preservation					Prompt Following				CP·PF
		`Animal`	`Human`	`Object`	`Style`	Overall	`Realistic`	`Style`	`Imaginative`	Overall	
• DreamBooth LoRA	SDXL v1.0	0.751	0.311	0.543	0.718	0.598	0.898	0.895	0.754	0.865	0.517
• IP-Adapter ViT-G	SDXL v1.0	0.667	0.558	0.504	0.752	0.593	0.743	0.632	0.446	0.640	0.380
• Emu2	SDXL v1.0	0.670	0.546	0.447	0.454	0.528	0.732	0.719	0.560	0.690	0.364
• DreamBooth	SD v1.5	0.640	0.199	0.488	0.476	0.494	0.789	0.775	0.504	0.721	0.356
• IP-Adapter-Plus ViT-H	SDXL v1.0	0.900	0.845	0.759	0.912	0.833	0.502	0.384	0.279	0.413	0.344
• BLIP-Diffusion	SD v1.5	0.673	0.557	0.469	0.507	0.547	0.581	0.510	0.303	0.495	0.271
• Textual Inversion	SD v1.5	0.502	0.358	0.305	0.358	0.378	0.671	0.686	0.437	0.624	0.236

3.2 Main Results

Quantitative & Qualitative Analysis Table 1 shows the overall evaluation results, including human and GPT-4o rating scores. The results show that: **i) DreamBench++ *aligns better with humans than DINO or CLIP models***. Driven by our dedicatedly-designed prompts, GPT-4o used by DreamBench++ yields impressive alignment with humans. This is because humans and DreamBench++ are all advanced in evaluating facial and textural characters and producing scores with a balanced consideration. **ii)** DINO-I and CLIP-I yield significant divergence from humans in evaluating concept preservation. This could be because DINO/CLIP scores show a preference for images that preserve shapes or overall styles (see Fig. 6). **iii)** Traditional CLIP-T scores are as effective as DreamBench++ in evaluating prompt following, showing strong alignment with humans. See qualitative results in Appendix C for an intuitive understanding of evaluated models.

Leaderboard Table 2 shows the leaderboard results with respect to the concept and prompt categories defined in Section 2. Note that: **i)** the `human` category shows the lowest average score of 0.482, which is -0.204 lower than the highest average score of `animal`. This category is very challenging in terms of concept preservation because due to facial details, and many works are conducted specifically on it (Wang et al., 2024b; Xiao et al., 2023; Valevski et al., 2023; Yan et al., 2023). **ii)** The `object` is also a relatively difficult category due to object diversity. In contrast, animals within the same category often share a strong visual similarity. **iii)** There exists a negative correlation between concept preservation and prompt following. The primary aim of personalized T2I evolution is to identify the Pareto optimum that balances both factors.

3.3 Ablation Study

Table 3 shows the ablation study of the prompt design influences on alignment. We observe that: **i)** The proposed prompt designs are all necessarily effective, demonstrating the superiority of the prompting method in DreamBench++. For example, removing the proposed internal thinking leads to a significant drop, indicating the effectiveness of self-alignment. **ii)** The capability of the multimodal GPT used is scalable. This shows that DreamBench++ has the potential to be improved in the future. **iii)** Some human prior knowledge, such as reminding the GPT not to consider background when assessing visual concept preservation, leads to performance degradation.

Table 3: **Ablation study of prompt design.** H, G, D, and C represent Human, GPT-4o, DINO Score, and CLIP Score, respectively. H-H value is also calculated to illustrate human self-alignment.

Method **T2I Model**	TI SD v1.5	DreamBooth SD v1.5	DreamBooth-L SDXL v1.0	BLIP-D SD v1.5	Emu2 SDXL v1.0	IP-Adapt.-P SDXL v1.0	IP-Adapt. SDXL v1.0
Concept Preservation							
H-H	0.685	0.647	0.656	0.613	0.746	0.602	0.591
G-H	0.544±0.014	0.596±0.003	0.641±0.007	0.362±0.017	0.669±0.005	0.366±0.017	0.458±0.002
- Internal Thinking	-0.040	-0.023	-0.012	+0.001	-0.045	-0.038	-0.008
- Scoring Criteria	-0.125	-0.116	-0.093	-0.158	-0.103	-0.227	-0.166
- Scoring Range	-0.038	-0.017	-0.027	-0.006	-0.016	-0.009	-0.017
+ Human Prior	-0.033	-0.022	-0.006	-0.015	-0.022	+0.009	-0.019
+ GPT4V	-0.105	-0.067	-0.131	-0.180	-0.016	-0.301	-0.250
Prompt Following							
H-H	0.475	0.516	0.469	0.619	0.441	0.576	0.509
G-H	0.461±0.007	0.506±0.002	0.402±0.001	0.541±0.003	0.422±0.011	0.484±0.006	0.531±0.006
- Internal Thinking	-0.013	+0.004	-0.032	-0.002	-0.014	+0.012	-0.002
- Scoring Criteria	-0.025	-0.012	-0.009	-0.012	-0.018	-0.017	-0.013
- Scoring Range	-0.010	-0.013	-0.011	+0.043	-0.038	+0.060	+0.036
+ GPT4V	-0.010	+0.012	0.000	-0.111	-0.007	-0.161	-0.134

4 Discussions

4.1 Is DreamBench++ aligned with humans?

Table 4 shows a more rigorous study of human alignment level using the mean Krippendorff's alpha value (Hayes & Krippendorff, 2007). The results show that **DreamBench++ is a highly human-aligned benchmark**. Notably, DreamBench++ achieves **79.64%** and **93.18%** evaluation consistency with human's evaluation in concept preservation and prompt following capabilities, respectively. This result is **+54.1%** and **+50.7%** higher than traditional DINO and CLIP metrics.

Table 4: **The human alignment degree among different evaluation metrics**, measured by Krippendorff's alpha value. H, G, D, and C represent Human, GPT-4o, DINO Score, and CLIP Score, respectively. H-H value is also calculated to illustrate human self-alignment.

Method	T2I Model	Concept Preservation $\mathrm{Kd}_{\bar{\alpha}}$				Prompt Following $\mathrm{Kd}_{\bar{\alpha}}$		
		H-H	G-H	D-H	C-H	H-H	G-H	C-H
• Textual Inversion	SD v1.5	0.685	0.544±0.014	0.262	-0.030	0.475	0.461±0.007	0.267
• DreamBooth	SD v1.5	0.647	0.596±0.003	0.408	0.229	0.516	0.506±0.002	0.185
• DreamBooth LoRA	SDXL v1.0	0.656	0.641±0.007	0.371	0.321	0.469	0.402±0.001	0.022
• BLIP-Diffusion	SD v1.5	0.613	0.362±0.017	-0.078	-0.186	0.619	0.541±0.003	0.319
• Emu2	SDXL v1.0	0.746	0.669±0.005	0.518	0.258	0.441	0.422±0.011	0.230
• IP-Adapter-Plus ViT-H	SDXL v1.0	0.602	0.366±0.017	-0.141	-0.150	0.576	0.484±0.006	0.256
• IP-Adapter ViT-G	SDXL v1.0	0.591	0.458±0.002	-0.073	-0.212	0.509	0.531±0.006	0.196
$\mathrm{Ratio}_{\bar{\alpha}}$		100%	79.64%	25.54%	3.34%	100%	93.18%	42.48%

4.2 Is data diversity necessary?

To assess the importance of diverse data, we compare results on DreamBench and DreamBench++ using DINO and CLIP metrics. Table 5 shows that **the diverse data in DreamBench++ is key to unbiased evaluation.** While overall results are consistent, TI, DreamBooth, and Emu2 show notable score drops. These methods perform well on natural images and simple text but struggle with complex or stylized prompts and anime references, see Fig. 7.

Table 5: **DreamBench and DreamBench++ results comparison with traditional metrics.** *Unlike DreamBench, DreamBench++ uses a single reference image per instance; thus, the training steps and learning rate of • fine-tuning-based methods are slightly reduced to avoid overfitting.

Method	T2I Model	DreamBench			DreamBench++		
		DINO-I	CLIP-I	CLIP-T	DINO-I	CLIP-I	CLIP-T
• Textual Inversion*	SD v1.5	0.557	0.753	0.259	0.437	0.726	0.302
• DreamBooth*	SD v1.5	0.678	0.786	0.301	0.544	0.753	0.323
• DreamBooth LoRA*	SDXL v1.0	0.646	0.769	0.325	0.628	0.784	0.341
• BLIP-Diffusion	SD v1.5	0.630	0.784	0.293	0.649	0.823	0.286
• Emu2	SDXL v1.0	0.753	0.842	0.283	0.539	0.763	0.310
• IP-Adapter-Plus ViT-H	SDXL v1.0	0.846	0.902	0.272	0.834	0.917	0.282
• IP-Adapter ViT-G	SDXL v1.0	0.681	0.835	0.295	0.667	0.855	0.309

Figure 7: **Case study of successful and failure case on DreamBench++.** The left images are reference images and the right images are results generated by Emu2, DreamBooth, and TI.

Table 6: **Study of Chain-of-Though (CoT) and In-context Learning (ICL) on human alignment.**

Method T2I Model	TI SD v1.5	DreamBooth SD v1.5	DreamBooth-L SDXL v1.0	BLIP-D SD v1.5	Emu2 SDXL v1.0	IP-Adapt.-P SDXL v1.0	IP-Adapt. SDXL v1.0
H-H	0.685	0.647	0.656	0.613	0.746	0.602	0.591
w/o CoT & w/o ICL	0.544	0.596	0.641	0.362	0.669	0.366	0.458
+ 1 shot ICL	-0.046	-0.019	-0.043	+0.013	-0.028	-0.098	-0.066
+ 2 shot ICL	-0.042	-0.023	-0.022	-0.033	-0.036	-0.085	-0.054
w/ CoT & w/o ICL	0.510	0.576	0.602	0.329	0.644	0.359	0.418
+ 1 shot ICL	-0.040	-0.008	-0.009	-0.020	-0.035	-0.145	-0.086
+ 2 shot ICL	-0.030	-0.002	-0.002	-0.051	-0.031	-0.155	-0.082

4.3 Can we use free lunch to improve DreamBench++ evaluation?

Table 6 shows the result of utilizing free lunch techniques, including chain-of-thought (CoT) (Wei et al., 2022) and In-Context Learning (ICL) (Alayrac et al., 2022; Brown et al., 2020). CoT indicates that GPT-4 articulates its reasoning process before scoring, and ICL indicates GPT-4o is provided with human-written few-shot examples.

Chain-of-Thought: i) CoT is effective in evaluating prompt following capability. Through CoT, the model more accurately discerns the significance of phrases such as "morphs into a mythical dragon", allowing it to assign a more appropriate evaluation score. ii) CoT does not bring improvement in concept preservation evaluation. We argue that CoT may shift attention to unnecessarily important background or texture information, as shown in Fig. 8.

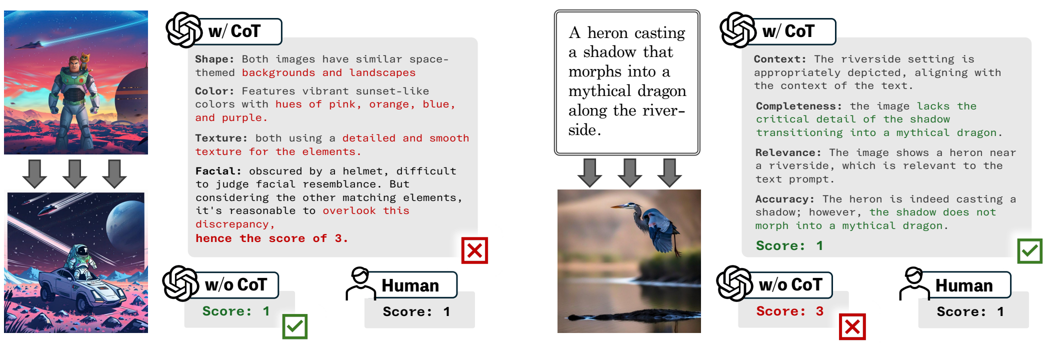


Figure 8: **Case study on CoT Prompting**. We find that (a) CoT prompting can improve text following evaluation by recognizing important parts of the prompt. (b) However, it may also hinder visual concept preservation by drifting GPT's attention away from the subject.

In-Context Learning: ICL counterintuitively leads to a drop in alignment. This could be attributed to the patching scheme, sample selection, or inherent bias within GPT-4o, making it non-trivial to prompt effectively. Thus, we provide our detailed prompt and hope to inspire future works.

4.4 Are multiple images for each instance necessary?

In practice, multiple reference images are unnecessary for personalized image generation: **i)** The limited availability of reference images during daily usage makes single-image personalization more relevant. **ii)** Fine-tuning methods perform well with just one image, as shown in Appendix C.

5 Conclusions

This paper introduces DreamBench++, a human-aligned personalized image generation benchmark. Extensive and comprehensive experiments show significant advantages in dataset diversity and complexity, along with metrics that align with human preferences. In addition, we offer insights into prompt design for advanced multimodal GPTs, emphasizing the potential and challenges of enhancing GPT evaluation through chain-of-thought prompting and in-context learning. Our work aims to support future research on personalized image generation by providing a human-aligned benchmark and heuristics in utilizing advanced multimodal GPTs in visual evaluation.

References

Google images. https://images.google.com. 5

Rawpixel. https://www.rawpixel.com. 5

Unsplash. https://unsplash.com. 5

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 9

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023. 19

Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In *SIGGRAPH Asia 2023 Conference Papers*, 2023. 1, 2

Moab Arar, Andrey Voynov, Amir Hertz, Omri Avrahami, Shlomi Fruchter, Yael Pritch, Daniel Cohen-Or, and Ariel Shamir. Palp: Prompt aligned personalization of text-to-image models. *CoRR*, abs/2401.06105, 2024. 2

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *CoRR*, abs/2211.01324, 2022. 1

James Betker, Goh Gabriel, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. *Computer Science. https://cdn.openai.com/papers/dall-e-3.pdf*, 2023. 1, 2

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023. 20

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Annu. Conf. Robot. Learn. (CoRL)*, 2023. 20

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 19

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2020. 9, 19

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Int. Conf. Comput. Vis. (ICCV)*, 2021. 2, 19

Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, José Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *Int. Conf. Mach. Learn. (ICML)*, 2023. 1

Hong Chen, Yipeng Zhang, Simin Wu, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In *Int. Conf. Learn. Represent. (ICLR)*, 2023a. 1

Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. Controlstyle: Text-driven stylized image generation using diffusion priors. In *ACM Int. Conf. Multimedia (ACM MM)*, pp. 7540–7548. ACM, 2023b. 1, 19

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W. Cohen. Subject-driven text-to-image generation via apprenticeship learning. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023c. 1, 2, 5

Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for step-by-step text-to-image generation and evaluation. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 19

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *Int. Conf. Learn. Represent. (ICLR)*, 2024. 19

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Adv. Neural Inform. Process. Syst. (NIPS)*, 2017. 4

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2021. 1

Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 1

Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 19

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. DreamLLM: Synergistic multimodal comprehension and creation. In *Int. Conf. Learn. Represent. (ICLR)*, 2024. 1, 2, 19

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *Int. Conf. Mach. Learn. (ICML)*, 2023. 20

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 19

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 1, 20

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Int. Conf. Learn. Represent. (ICLR)*, 2023a. 1, 2, 6, 19

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Trans. Graph.*, 42(4):150:1–150:13, 2023b. 1, 2, 19

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. SEED-X: multimodal models with unified multi-granularity comprehension and generation. *CoRR*, abs/2404.14396, 2024. 19

Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 20

Jianshu Guo, Wenhao Chai, Jie Deng, Hsiang-Wei Huang, Tian Ye, Yichen Xu, Jiawei Zhang, Jenq-Neng Hwang, and Gaoang Wang. Versat2i: Improving text-to-image models with versatile reward. *CoRR*, abs/2403.18493, 2024. 19

Chunrui Han, Jinrong Yang, Jianjian Sun, Zheng Ge, Runpei Dong, Hongyu Zhou, Weixin Mao, Yuang Peng, and Xiangyu Zhang. Exploring recurrent long-term temporal fusion for multi-view 3d perception. *IEEE Robotics and Automation Letters*, 9(7):6544–6551, 2024. 20

Andrew F Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007. 8

Junjie He, Yifeng Geng, and Liefeng Bo. Uniportrait: A unified framework for identity-preserving single-and multi-human image personalization. *CoRR*, abs/2408.05939, 2024a. 2

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 9726–9735. Computer Vision Foundation / IEEE, 2020. 19

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Annu. Conf. Robot. Learn. (CoRL)*, 2024b. 20

Zecheng He, Bo Sun, Felix Juefei-Xu, Haoyu Ma, Ankit Ramchandani, Vincent Cheung, Siddharth Shah, Anmol Kalia, Harihar Subramanyam, Alireza Zareian, et al. Imagine yourself: Tuning-free personalized image generation. *CoRR*, abs/2409.13346, 2024c. 2

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *Int. Conf. Learn. Represent. (ICLR)*, 2023a. 19

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. *CoRR*, abs/2312.02133, 2023b. 19

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Empir. Method. Nat. Lang. Process. (EMNLP)*, 2021. 19

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Adv. Neural Inform. Process. Syst. (NIPS)*, 2017. 19

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2020. 19

Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 20

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Int. Conf. Learn. Represent. (ICLR)*, 2022. 6, 19

Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. Instruct-imagen: Image generation with multi-modal instruction. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 4754–4763, 2024. 2

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Int. Conf. Comput. Vis. (ICCV)*, 2023. 19

Miao Hua, Jiawei Liu, Fei Ding, Wei Liu, Jie Wu, and Qian He. Dreamtuner: Single image is enough for subject-driven generation. *CoRR*, abs/2312.13691, 2023. 1

Jiannan Huang, Jun Hao Liew, Hanshu Yan, Yuyang Yin, Yao Zhao, and Yunchao Wei. Classdiffusion: More aligned personalization tuning with explicit class guidance. *CoRR*, abs/2405.17532, 2024a. 2

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 19

Mengqi Huang, Zhendong Mao, Mingcong Liu, Qian He, and Yongdong Zhang. Realcustom: Narrowing real text word for real-time open-domain text-to-image customization. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024b. 2

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017. 19

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Int. Conf. Mach. Learn. (ICML)*, 2021. 5

Xuhui Jia, Yang Zhao, Kelvin C. K. Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *CoRR*, abs/2304.02642, 2023. 1, 19

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 1

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 5

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *CoRR*, abs/2312.14867, 2023. 4, 19

Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. In *Int. Conf. Learn. Represent. (ICLR)*, 2024. 2, 3, 19

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 1, 2, 3, 5, 19

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Fei-Fei Li, Jiajun Wu, Stefano Ermon, and Percy Liang. Holistic evaluation of text-to-image models. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 2, 3, 19

Dongxu Li, Junnan Li, and Steven C. H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023a. 1, 6, 19

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Int. Conf. Mach. Learn. (ICML)*, 2022. 19

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Int. Conf. Mach. Learn. (ICML)*, 2023b. 19

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1477–1485, 2023c. 20

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023a. 20

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J. Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation. *CoRR*, abs/2312.10240, 2023b. 19

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744, 2023a. 19

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023b. 19

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Int. Conf. Comput. Vis. (ICCV)*, 2023c. 20

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. In *Int. Conf. Mach. Learn. (ICML)*, 2023d. 19

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 19

Henglei Lv, Jiayu Xiao, Liang Li, and Qingming Huang. Pick-and-draw: Training-free semantic guidance for text-to-image personalization. *CoRR*, abs/2401.16762, 2024. 1, 19

Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *CoRR*, abs/2303.09319, 2023. 1

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Int. Conf. Mach. Learn. (ICML)*, 2021. 19

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *Int. Conf. Mach. Learn. (ICML)*, 2022. 1

OpenAI. Gpt-4v(ision) system card. 2023. URL `https://openai.com/research/gpt-4v-system-card`. 4, 19

OpenAI. Introducing gpt-4o and more tools to chatgpt free users. 2024. URL `https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/`. 2, 4

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 4

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. In *Int. Conf. Learn. Represent. (ICLR)*, 2024. 2, 19

Lianyu Pang, Jian Yin, Haoran Xie, Qiping Wang, Qing Li, and Xudong Mao. Cross initialization for personalized text-to-image generation. 2024a. 2

Lianyu Pang, Jian Yin, Baoquan Zhao, Feize Wu, Fu Lee Wang, Qing Li, and Xudong Mao. Attndreambooth: Towards text-aligned personalized text-to-image generation. *CoRR*, abs/2406.05000, 2024b. 2

Jae Wan Park, Sang Hyun Park, Jun Young Koh, Junha Lee, and Min Song. CAT: contrastive adapter training for personalized image generation. *CoRR*, abs/2404.07554, 2024. 19

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, August 6-10, 2023*, pp. 11:1–11:11, 2023. 19

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *Int. Conf. Learn. Represent. (ICLR)*, 2024. 6

Pavlin G. Poličar, Martin Stražar, and Blaž Zupan. opentsne: a modular python library for t-sne dimensionality reduction and embedding. *bioRxiv*, 2019. URL https://github.com/pavlin-policar/openTSNE. 5, 6

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 20

Zekun Qi, Muzhou Yu, Runpei Dong, and Kaisheng Ma. VPP: efficient conditional 3d generation via voxel-point progressive representation. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 20

Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *Computer Vision – ECCV 2024*, pp. 214–238, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72775-7. 20

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. 36:79320–79362, 2023. 2

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn. (ICML)*, 2021. 2, 19

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Int. Conf. Mach. Learn. (ICML)*, 2021. 1

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. 1, 2

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530, 2024. 19

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 1, 6

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 1, 2, 3, 5, 6, 19

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH, 2022*, pp. 15:1–15:10. ACM, 2022a. 19

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022b. 1

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Adv. Neural Inform. Process. Syst. (NIPS)*, 2016. 19

Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, number FZJ-2022-00923. Jülich Supercomputing Center, 2021. 5

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 20

Kihyuk Sohn, Lu Jiang, Jarred Barber, Kimin Lee, Nataniel Ruiz, Dilip Krishnan, Huiwen Chang, Yuanzhen Li, Irfan Essa, Michael Rubinstein, Yuan Hao, Glenn Entis, Irina Blok, and Daniel Castro Chin. Styledrop: Text-to-image synthesis of any style. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 19

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Int. Conf. Learn. Represent. (ICLR)*, 2021a. 19

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Int. Conf. Learn. Represent. (ICLR)*, 2021b. 19

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024a. 2, 6, 19

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *Int. Conf. Learn. Represent. (ICLR)*, 2024b. 2

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David D. Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 2, 4

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, August 6-10, 2023*, pp. 12:1–12:11. ACM, 2023. 1

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 1, 19

Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. In *SIGGRAPH Asia 2023 Conference Papers, SA 2023, Sydney, NSW, Australia, December 12-15, 2023*, pp. 94:1–94:10. ACM, 2023. 3, 7, 19

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *J. Mach. Learn. Res. (JMLR)*, 9(11), 2008. 5, 6

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Int. Conf. Mach. Learn. (ICML)*, volume 307 of *ACM International Conference Proceeding Series*, pp. 1096–1103. ACM, 2008. 19

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *CoRR*, abs/2404.02733, 2024a. 1, 19

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *CoRR*, abs/2305.17926, 2023a. 4

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *CoRR*, abs/2401.07519, 2024b. 1, 3, 7, 19

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018. 19

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Association for Computational Linguistics*, 2023b. 4

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. *CoRR*, abs/2407.05600, 2024c. 2

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 2, 5, 9

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: encoding visual concepts into textual embeddings for customized text-to-image generation. In *Int. Conf. Comput. Vis. (ICCV)*, 2023. 1, 19

Feize Wu, Yun Pang, Junyi Zhang, Lianyu Pang, Jian Yin, Baoquan Zhao, Qing Li, and Xudong Mao. Core: Context-regularized text embedding learning for text-to-image personalization. *CoRR*, abs/2408.15914, 2024a. 2

Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas J. Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024b. 4, 19

Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *CoRR*, abs/2305.10431, 2023. 1, 3, 7, 19

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *CoRR*, abs/2409.11340, 2024. 2

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 2, 19

Paul Yacoubian. Avocado bag, July 2022. URL https://twitter.com/PaulYacoubian/status/1542867718071779330. 2

Yuxuan Yan, Chi Zhang, Rui Wang, Yichao Zhou, Gege Zhang, Pei Cheng, Gang Yu, and Bin Fu. Facestudio: Put your face everywhere in seconds. *CoRR*, abs/2312.02663, 2023. 3, 7, 19

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *CoRR*, abs/2308.06721, 2023. 1, 2, 3, 6, 19

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *T. Mach. Learn. Res. (TMLR)*, 2022. 1, 5

Linfeng Zhang, Xin Chen, Junbo Zhang, Runpei Dong, and Kaisheng Ma. Contrastive deep supervision. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 19

Linfeng Zhang, Xin Chen, Runpei Dong, and Kaisheng Ma. Region-aware knowledge distillation for efficient image-to-image translation. In *Brit. Mach. Vis. Conf. (BMVC)*, 2023a. 19

Linfeng Zhang, Runpei Dong, Hung-Shuo Tai, and Kaisheng Ma. Pointdistiller: structured knowledge distillation towards efficient and compact 3d detection. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023b. 20

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018. 19

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v(ision) as a generalist evaluator for vision-language tasks. *CoRR*, abs/2311.01361, 2023c. 4, 19

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. In *Int. Conf. Mach. Learn. (ICML)*, 2023d. 2, 5

Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, and Xiangyu Zhang. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. In *Int. Joint Conf. Artif. Intell. (IJCAI)*, 2024. 20

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 4

Matthew Zheng, Enis Simsar, Hidir Yesiltepe, Federico Tombari, Joel Simon, and Pinar Yanardag. Stylebreeder: Exploring and democratizing artistic styles through text-to-image models. *CoRR*, abs/2406.14599, 2024. 19

Yufan Zhou, Ruiyi Zhang, Kaizhi Zheng, Nanxuan Zhao, Jiuxiang Gu, Zichao Wang, Xin Eric Wang, and Tong Sun. Toffee: Efficient million-scale dataset construction for subject-driven text-to-image generation. *CoRR*, abs/2406.09305, 2024. 1

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Int. Conf. Comput. Vis. (ICCV)*, 2017. 19

A Related Works

Personalized Image Generation aims to preserve concept consistency while accommodating the diverse contexts suggested by the instructions. In general, it can be traced back to early efforts on pixel-to-pixel (Pix2Pix) translation where the personalization orientation is free-form texts (Brooks et al., 2023; Tumanyan et al., 2023; Parmar et al., 2023) or predefined translation across styles, seasons, species, or plants, *etc* (Isola et al., 2017; Zhu et al., 2017; Zhang et al., 2023a; Wang et al., 2018; Saharia et al., 2022a). Modern efforts go beyond Pix2Pix translation toward a free-form image generation conditioned on both reference images and prompts. Some works focus on fine-tuning techniques that turn a general T2I model into a specialist personalization model (Gal et al., 2023a; Ruiz et al., 2023; Kumari et al., 2023; Sohn et al., 2023; Park et al., 2024) using LoRA (Hu et al., 2022) or contrastive learning (Zhang et al., 2022; He et al., 2020), learning the subject or style information by reconstructive autoencoding (Vincent et al., 2008; Dong et al., 2023). However, the necessity to fine-tune for new subjects limits their scalability. In contrast, encoder-based methods can generate subject-guided or style-guided images or edit images following prompts with one shot. Encoder- or adapter-based methods (Zheng et al., 2024; Ye et al., 2023; Wei et al., 2023; Li et al., 2023a; Jia et al., 2023; Gal et al., 2023b; Chen et al., 2023b; Wang et al., 2024a;b) train an encoder to encode the conditional image into embeddings, which are integrated into cross-attention mechanism in the diffusion process (Ho et al., 2020; Song et al., 2021a; Nichol & Dhariwal, 2021; Song et al., 2021b). Adapter-free methods (Lv et al., 2024; Liu et al., 2023d; Hertz et al., 2023b;a; Brooks et al., 2023) extract the information, such as attention maps (Hertz et al., 2023a) from reference images, and fuse them into the image generation process. Furthermore, multimodal large language models (MLLMs) that are trained on extensive multimodal sequences can also serve as general foundation models (Dong et al., 2024; Sun et al., 2024a; Pan et al., 2024; Ge et al., 2024). Additionally, some works (Wang et al., 2024b; Valevski et al., 2023; Yan et al., 2023; Ye et al., 2023; Xiao et al., 2023) focus on facial feature preservation.

Benchmarking Image Generation involves a variety of metrics that focus on different aspects. Inception Score (Salimans et al., 2016) and FID (Heusel et al., 2017) judge image quality, while LPIPS (Zhang et al., 2018), DreamSim (Fu et al., 2023), CLIP-I (Radford et al., 2021), and DINO Score (Caron et al., 2021) measure perceptual similarity. In text-guided generation, prompt-image alignment can be assessed by CLIP-T (Radford et al., 2021), CLIPScore (Hessel et al., 2021), and BLIP Score (Li et al., 2022; 2023b). However, these metrics often fall short of reflecting human perception. To address this, human-aligned metrics (Ku et al., 2024; Xu et al., 2023; Lee et al., 2023) have been introduced, offering a more perceptive evaluation. Yet, they face limitations in scaling with the pace of new model developments. Thus, the necessity for automated and sustainable evaluation methods has emerged, with some (Xu et al., 2023; Liang et al., 2023b; Guo et al., 2024) leveraging reward-model-based methods to encode human preferences, while others (Ku et al., 2023; Cho et al., 2023; Wu et al., 2024b; Zhang et al., 2023c; Hu et al., 2023; Lu et al., 2023) use multimodal (Brown et al., 2020; Reid et al., 2024; Anil et al., 2023; Liu et al., 2023b) to automate the process and better mirror human tastes. While MLLM-based methods show promise in aligning with human preferences (Zhang et al., 2023c; Wu et al., 2024b; Huang et al., 2023; Cho et al., 2024), automated personalized evaluation remains an unresolved issue. VIEScore (Ku et al., 2023) assesses image generation quality by prompting GPT-4V (OpenAI, 2023) and LLaVA (Liu et al., 2023b;a), but is limited to four models in subject-driven tasks and obtains suboptimal results. Meanwhile, Dreambench (Ruiz et al., 2023), a common benchmark for personalized generative evaluation, only consists of 30 simple objects and lacks diversity comprehensiveness.

B Implementation Details

The configurations for the training hyperparameters used in training-based methods on DreamBench and DreamBench++, are detailed in Table 7. During the inference stage, all methods employ a `guidance_scale` of 7.5 and execute 100 inference steps, with the exception of Emu2, which uses a `guidance_scale` of 3 and performs 50 inference steps. Furthermore, BLIP-Diffusion and IP-Adapter incorporate negative prompts, as demonstrated in Table 8. Specifically, IP-Adapter includes an additional parameter, `ip_adapter_scale`, set at 0.6.

Table 7: **Training hyperparameters on DreamBench and DREAMBENCH++**. BS: batch size, LR: learning rate, Steps: training steps.

		DreamBench			DREAMBENCH++		
Method	**T2I Model**	**BS**	**LR**	**Steps**	**BS**	**LR**	**Steps**
Textual Inversion	SD v1.5	4	5e-4	3000	1	5e-4	3000
Dreambooth	SD v1.5	1	2.5e-6	1000	1	2.5e-6	250
Dreambooth LoRA	SDXL v1.0	4	5e-5	500	1	5e-5	500

Table 8: **Negative Prompt Templates**

Method	**T2I Model**	**Negative Prompt**
BLIP-Diffusion	SD v1.5	over-exposure, under-exposure, saturated, duplicate, out of frame, lowres, cropped, worst quality, low quality, jpeg artifacts, morbid, mutilated, ugly, bad anatomy, bad proportions, deformed, blurry, duplicate
IP-Adapter ViT-G	SDXL v1.0	deformed, ugly, wrong proportion, low res, bad anatomy, worst quality, low quality
IP-Adapter-Plus ViT-H	SDXL v1.0	deformed, ugly, wrong proportion, low res, bad anatomy, worst quality, low quality

We dedicate significant effort to tuning hyper-parameters to ensure that the performance of each method on DreamBench is consistent with results reported in original papers. Table 9 shows the results of our reproduction are comparable to or even better than the official results.

Table 9: **Reproduced results**. Our reproduction is comparable to or better than the official results. N/A denotes that the official paper does not report the corresponding results.

Method	**T2I Model**	**DINO-I**		**CLIP-I**		**CLIP-T**	
		Official	**Reproduction**	**Official**	**Reproduction**	**Official**	**Reproduction**
• Textual Inversion	SD v1.5	0.569	0.557	0.780	0.753	0.255	0.259
• DreamBooth	SD v1.5	0.688	0.678	0.803	0.786	0.305	0.301
• DreamBooth LoRA	SDXL v1.0	N/A	0.646	N/A	0.769	N/A	0.325
• BLIP-Diffusion	SD v1.5	0.594	0.630	0.779	0.784	0.300	0.293
• Emu2	SDXL v1.0	0.766	0.753	0.850	0.842	0.287	0.283
• IP-Adapter-Plus ViT-H	SDXL v1.0	N/A	0.846	N/A	0.902	N/A	0.272
• IP-Adapter ViT-G	SDXL v1.0	N/A	0.681	N/A	0.835	N/A	0.295

C QUALITATIVE ANALYSIS

With a more comprehensive and diverse collection of images, we have discovered numerous intriguing characteristics of these generation methods, as illustrated in Fig. 9, that were not apparent on existing datasets such as DreamBench. Specifically, we observe that: **i)** Fine-tuning-based methods outperform encoder-based methods on images containing more *subject-oriented* information, such as an animal or object, as they preserve more intricate details in the generated images. However, for images containing a person, fine-tuning-based methods often fail to preserve facial and clothing features. This suggests that the personalized generation of human images is more demanding for visual concept preservation than for textual following capabilities, which is an advantage for encoder-based methods. **ii)** However, for *style-oriented* cases when subject details are less critical, encoder-based methods perform better than fine-tuning-based methods. This further highlights the strengths of encoder-based methods in that they are more adept at recognizing and extracting high-level visual semantics, including overall shape, style, and thematic features.

D LIMITATION & FUTURE WORK

Human-aligned evaluation & benchmarking is an emerging but challenging direction, and we have only made preliminary attempts at personalized image generation. Moreover, our evaluation results heavily rely on the advancements of multimodal large language models and require carefully designed system prompts. We believe that as visual world models continue to develop, the evaluation performance will be further optimized. Our future work will focus on more applications with human-aligned evaluation, such as 3D generation (Poole et al., 2023; Qi et al., 2023; Liu et al., 2023c; Gafni et al., 2022), video generation (Ho et al., 2022; Singer et al., 2023; Blattmann et al., 2023), autonomous driving (Han et al., 2024; Li et al., 2023c; Zhang et al., 2023b), and even embodied visual intelligence (Goyal et al., 2022; Liang et al., 2023a; Brohan et al., 2023; Driess et al., 2023; Qi et al., 2025; Zhao et al., 2024; He et al., 2024b).

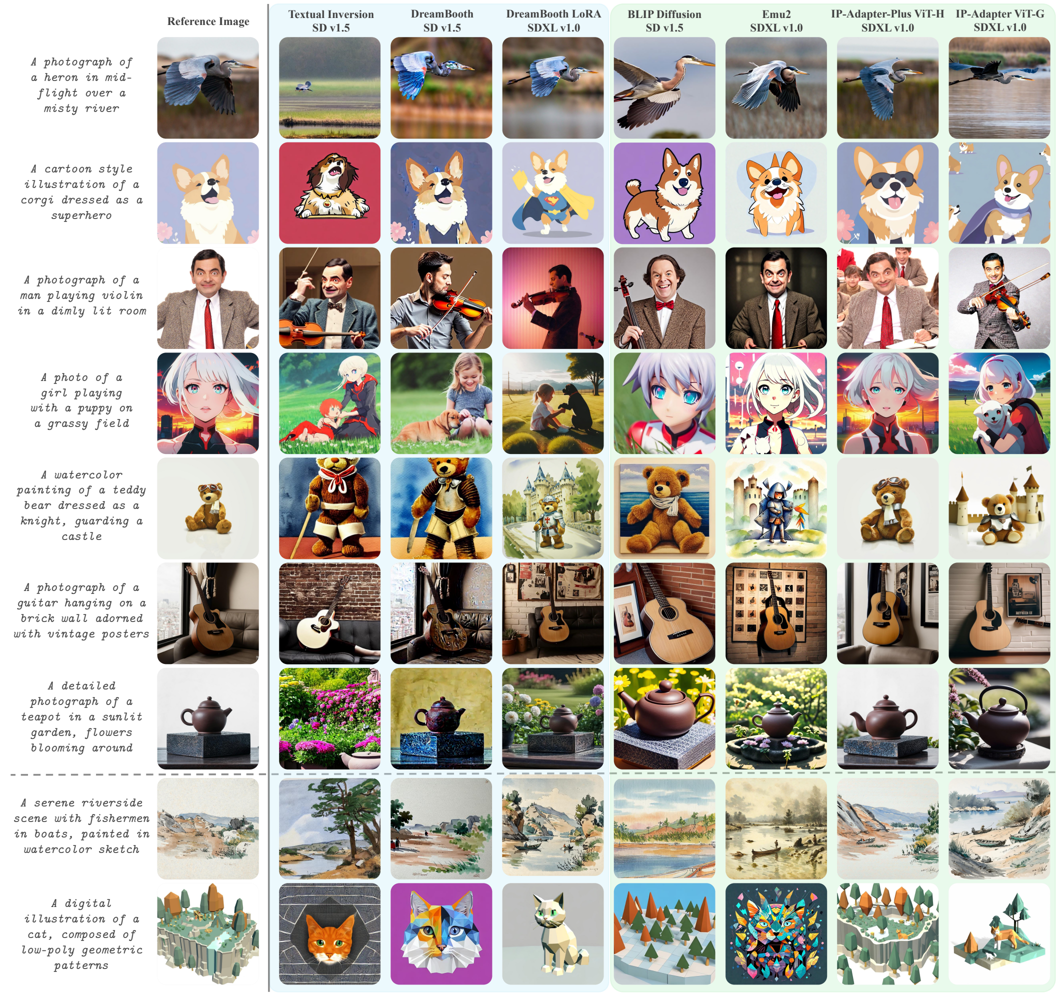


Figure 9: **A qualitative study of different methods on DREAMBENCH++.** We demonstrate the generation quality of different methods on our DREAMBENCH++, including animals, humans, objects, and style, with photo and non-photo-realistic examples. The blue block highlights fine-tuning-based methods, and the green block highlights encoder-based methods. Instances above the dotted line are evaluated for subject preserving, and instances below are evaluated for style preserving.

BROADER IMPACT

Powerful as the T2I generative models pretrained on large-scale web-scraped data, the models may be misused as illegal or unethical tools for generating NSFW content. This potential impact can also be brought by personalized T2I models as they are typically built on the pretrained T2I foundation models. As a result, it is critical to use tools such as NSFW detectors to avoid such content during both usage and evaluation. For example, the data used for evaluation must avoid the NSFW content by data filtering. In this paper, such contents are filtered out by human annotators.