ShareGPT-4o-Image: Aligning Multimodal Models with *GPT-4o*-Level Image Generation

Anonymous ACL submission

Abstract

Recent advances in multimodal generative models have unlocked photorealistic, instruction-aligned image generation, yet leading systems like GPT-4o-Image remain proprietary and inaccessible. To democratize these capabilities, we present **ShareGPT-4o-Image**, a synthetic distillation dataset comprising 45K text-to-image pairs and 46K text-and-image-to-image triplets, all synthesized using GPT-4o's image generation capabilities as compact supervision for open multimodal generators. Leveraging this dataset, we develop **Janus-4o**, a multimodal large language model capable of both text-to-image and text-and-image-to-image generation. Rather than aiming to replace the strongest commercial image generators, our study shows that a compact, high-quality distillation dataset plus lightweight adaptation can continuously improve an existing multimodal generator: Janus-4o improves text-to-image generation over Janus-Pro and adds text-and-image-to-image generation to Janus-Pro. It obtains these gains using only 91K synthetic samples and 6 hours of training on a single machine with 8×A800 GPUs. We hope the release of ShareGPT-4o-Image and Janus-4o will foster open research in data-efficient adaptation for photorealistic, instruction-aligned image generation.

1 Introduction

Recently, cutting-edge proprietary models have demonstrated unprecedented capabilities in both text-to-image and text-and-image-to-image generation, such as GPT-4o (Open AI, 2025), Seedream (Seedream, 2025) and Nano Banana[1] (Comanici et al., 2025). These models showcase impressive instruction alignment, semantic consistency, and photorealism for image generation and editing. They truly unlock the productivity and creativity of image AI, making image synthesis and editing incredibly user-friendly (Chen et al., 2025b; Qian et al., 2025; Wang et al., 2025b).

Despite the success of these proprietary models, they provide limited visibility into their training data, architectures, and optimization recipes. The immense computational resources required for training also put them out of reach for the broader research community. This motivates data-efficient approaches for transferring such capabilities to open models under practical compute budgets (Ye et al., 2025a).

In this context, this work first proposes to construct a large-scale and diverse distillation dataset fully synthesized with proprietary models. The benefits of this approach include: 1) high-quality, semantically consistent images generated by proprietary models, 2) data diversity without reliance on real image collections, and 3) creative flexibility beyond scene limitations. We propose a scalable method for constructing such a distillation dataset using LLMs to synthesize multi-scenario prompts, which are then used to distill proprietary models. Based on this, we have created ShareGPT-4o-Image, a synthesized dataset designed to distill the image generation capabilities of GPT-4o. The dataset includes 45K text-to-image and 46K text-and-image-to-image prompts. The resulting dataset spans a wide range of styles and grounded visual reasoning, highlighting proprietary model's strengths in instruction-following and visual aesthetics.

To validate the effectiveness of the distillation dataset, we fine-tune Janus-Pro (Chen et al., 2025c) on ShareGPT-4o-Image to develop Janus-4o, a novel multimodal large language model (MLLM) that supports both text-to-image and text-and-image-to-image generation. In text-to-image generation, Janus-4o improves on Janus-Pro by 4 and 1.6 points on the GenEval and DPG-Bench benchmarks, respectively. Unlike its predecessor,

[1] Nano Banana refers to the image generation feature of Gemini 2.5 Flash, later also available in Gemini 3 Pro.

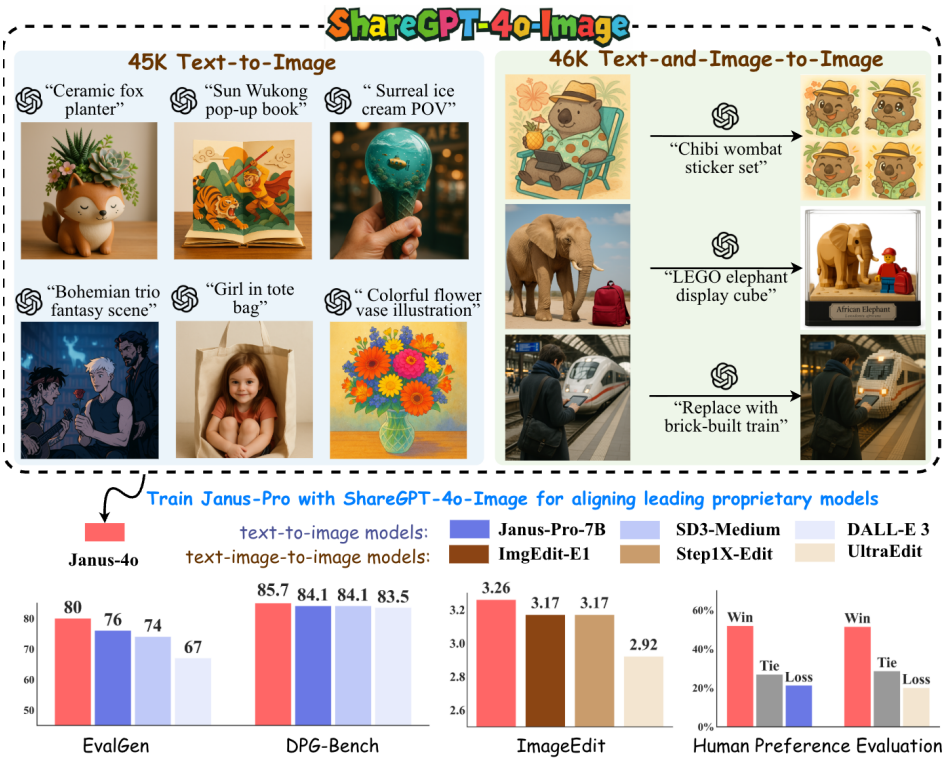


Figure 1: Overview of ShareGPT-4o-Image. ShareGPT-4o-Image is a 91K diverse image generation dataset distilled from GPT-4o, capturing the advanced capabilities of leading proprietary models for both *text-to-image* and *text-and-image-to-image* generation tasks. Fine-tuning Janus-Pro with ShareGPT-4o-Image yields **Janus-4o**, which shows notable improvements in image generation.

which only handled text-to-image tasks, Janus-4o introduces text-and-image-to-image capabilities through lightweight training adaptations. On ImgEdit-Bench, it remains competitive with models trained on far larger editing datasets, despite using only 91K samples. Human evaluations and ablation studies further show that high-quality distillation data can consistently enhance existing multimodal generators. We are releasing both the dataset and model to the research community to foster progress in data-efficient adaptation for open-source multimodal generation.

- We introduce ShareGPT-4o-Image, a large-scale distillation dataset containing *text-to-image* and *text-and-image-to-image* pairs for adapting open multimodal generators with supervision from proprietary models.
- Based on ShareGPT-4o-Image, we develop Janus-4o, a novel multimodal large language model capable of both *text-to-image* and *text-and-image-to-image* generation.
- Experiments show that Janus-4o significantly improves upon Janus-Pro in image generation and enables *text-and-image-to-image* generation. Remarkably, Janus-4o acquires this new capability using only 91K synthetic samples and 6 hours of training on a single 8×A800 GPU machine.

2 Related Work

2.1 Image Generation Models

Instruction-guided generation is mainly driven by diffusion models (Ho et al., 2020; Rombach et al., 2022; Ge et al., 2024b; Zhang et al., 2025; Cai et al., 2025) and autoregressive models (Ramesh et al., 2021; Esser et al., 2021; Wu et al., 2025; Team, 2024; Zhan et al., 2024). The former iteratively denoises random noise for high-fidelity image synthesis, while the latter discretizes images into token sequences, enabling controllable generation and integration with LLMs. Recent image editing works (Yu et al., 2024; Ye et al., 2025b; Zhao et al., 2024; Zhang et al., 2023; Brooks et al., 2022) further extend generation to instruction-driven modifications. *To the best of our knowledge, Janus-4o is among the first autoregressive models capable of both image generation and image editing.*

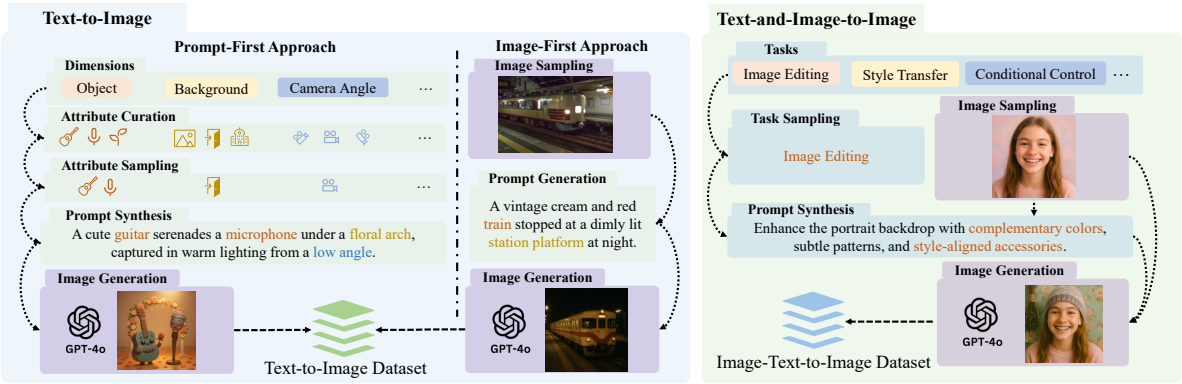


Figure 2: The flow diagram for the dataset construction process. The prompt in the diagram is the simplified version.

2.2 Datasets for Image Generation

Early image-text datasets (Lin et al., 2014; Sidorov et al., 2020; Changpinyo et al., 2021; Schuhmann et al., 2022) enabled *text-to-image* training at increasing scale, while image editing datasets (Brooks et al., 2022; Zhang et al., 2023; Zhao et al., 2024; Yu et al., 2024) support instruction-based visual modification. However, real-image targets can be noisy, low-quality, or inconsistent, and high-quality datasets for distilling advanced proprietary generators remain scarce. *To the best of our knowledge, our work is among the first to propose a distillation dataset with cutting-edge proprietary image generation models and demonstrate its effectiveness.* Subsequent works have followed our approach to construct distillation datasets (Qian et al., 2025; Ye et al., 2025a; Wang et al., 2025b).

2.3 Multimodal Large Language Models

LLMs have increasingly evolved toward multimodal interaction, with many MLLMs aligning visual features with text for image understanding (Achiam et al., 2023; Liu et al., 2023; Bai et al., 2025; Team et al., 2023; Chen et al., 2024d). Image generation is a natural extension: early systems used external tools (Fei et al., 2024), later approaches decoded images from LLM representations (Sun et al., 2024b; Wu et al., 2024; Zhan et al., 2024), and Janus-like models generate images autoregressively with image tokens (Tian et al., 2024; Wu et al., 2025). Such models provide a simple end-to-end design and leverage native LLM understanding for better instruction following. *Janus-4o is among the first models of this kind to support both text-to-image and text-and-image-to-image generation.*

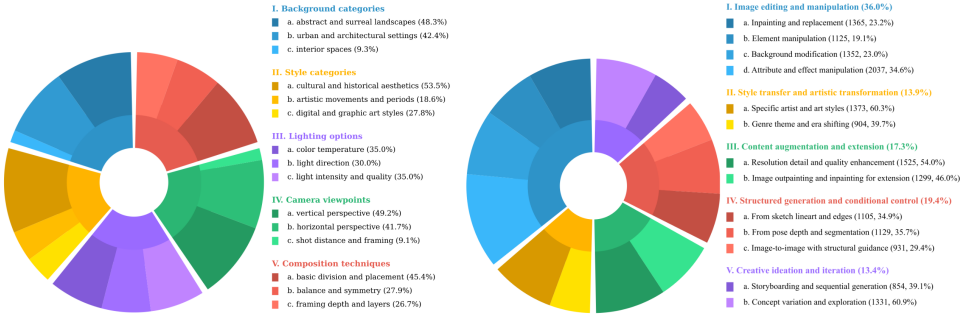
3 ShareGPT-4o-Image

To harness the advanced generative capabilities of leading proprietary image generation models like GPT-4o (Open AI, 2025), we propose ShareGPT-4o-Image, a large-scale dataset comprising 45K text-to-image pairs and 46K instruction-guided image editing triplets. The distillation process focuses on creating diverse and balanced input prompts that elicit the proprietary model's capabilities, rather than filtering prompts by the limitations of open-source target models. We describe the construction of text-to-image distillation data in Sec. 3.1 and text-and-image-to-image distillation in Sec. 3.2. The entire data generation process, illustrated in Figure 2, utilizes Gemini-Pro-2.5 (Team et al., 2023) as our primary Large Language Model (LLM) for all text synthesis tasks.

3.1 Text-to-Image Data

For text-to-image data, we need to construct diverse text prompts for distilling proprietary models. To maximize the diversity and complexity of the input prompt, we propose two complementary approaches: a **prompt-first** approach for controlled attribute coverage and an **image-first** approach to ground prompts in realistic visual content.

Prompt-First Pipeline. We start by defining a six-dimensional attribute space (Objects, Background, Style, Lighting, Camera, Composition) populated with a rich vocabulary, including 1,000 object categories from ImageNet (Appendix C). For each sample, we randomly sample attributes from these dimensions. The distribution used for sampling the number of object categories, k, is based on a random selection ranging from 1 to 100. We apply weights to ensure diverse sampling patterns, especially when multiple objects interact. We



(a) Text-to-Image Domain Distribution **(b) Text-and-Image-to-Image Task Distribution**

Figure 3: Data Distribution of ShareGPT-4o-Image. Figure (a) illustrates the selection proportions for each of the 5 dimensions in the text-to-image task, with the object dimension excluded due to its 1k items. Figure (b) presents the image editing task, categorized into 14 sub-categories and 5 primary categories. Detailed information can be found in Appendix C.

also employ an exponential decay distribution to model the likelihood of selecting a higher number of categories. This ensures that the probability of selecting more categories decreases exponentially, making it more likely to sample fewer categories, which is common in many scenarios.

The distribution is defined as follows:

$$w(x) = \exp(-\lambda(x-1))$$

where λ is the parameter controlling the decay rate, with a larger λ resulting in faster decay. Once patterns are sampled, we use Gemini-Pro-2.5 to compose them into a coherent, natural-language prompt, following the template in Appendix D. Finally, each prompt is passed to GPT-4o to generate a paired image. This systematic approach ensures the creation of diverse prompts with controlled complexity.

Image-First Pipeline. To complement the synthetic prompts, we source real-world, diverse, high-quality images from the ALLaVA dataset (Chen et al., 2024a). To ensure the diversity of the images, we first use the *clip-vit-large-patch14-336* (Radford et al., 2021) visual encoder to remove overly similar images based on image representation similarity, retaining 200K images for sampling. For each image, we prompt Gemini-Pro-2.5 to generate a detailed, descriptive text prompt that accurately reflects the visual content (see Appendix D.1 for the meta-prompt). Finally, the generated text prompt is input into GPT-4o to generate a paired image. The image-first approach enhances the diversity of the prompts, particularly by increasing the authenticity and complexity of the descriptions.

3.2 Text-and-Image-to-Image Data

To generate data for image editing, the text-and-image-to-image component of ShareGPT-4o-Image is constructed through a multi-stage process. This pipeline produces a diverse set of image modification instructions and their corresponding visual outcomes, as described below:

Task Framework Definition. We begin by defining a framework consisting of 14 sub-categories for the text-and-image-to-image tasks, which are organized into 5 primary categories. These tasks cover common editing operations, such as object replacement, style modification, and attribute alteration. A detailed overview of these categories is provided in Appendix C.1.1.

Input Assembly and Instruction Generation In this stage, a source image is selected, an editing task is sampled, and the LLM is used to generate a textual modification instruction. Each data point is associated with a unique source image, chosen from one of two pools: (1) images generated during the text-to-image data construction process, and (2) curated real-world photographs (ALLaVA dataset). Editing tasks are selected by first randomly choosing one of the 5 primary categories, followed by a selection of a sub-category from within that category. For each selected source image and editing task, the LLM generates a corresponding natural language instruction. The meta-prompts used to guide the LLM, which vary by task category, are detailed in Appendix D.

Image Synthesis Finally, the input image and editing instructions are fed into GPT-4o, which

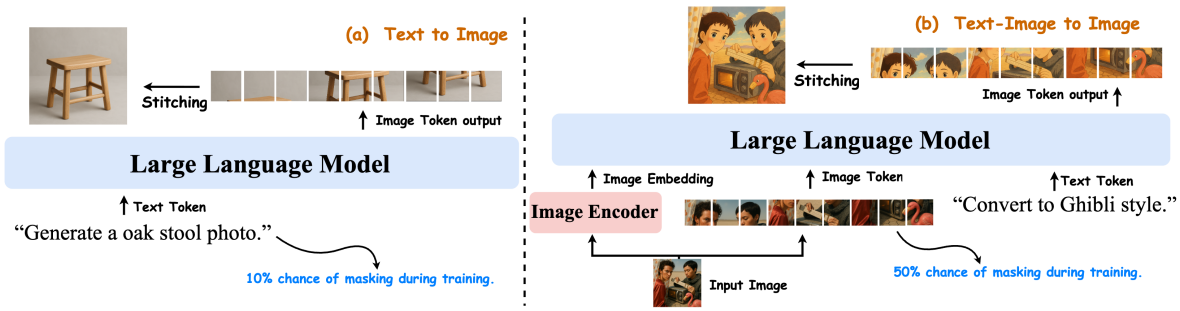


Figure 4: Overview of Janus-4o model. Built upon Janus-Pro, it is constructed via fine-tuning on ShareGPT-4o-Image. It incorporates enhancements to support *text-and-image-to-image* generation. Both *text-to-image* and *text-and-image-to-image* tasks are jointly trained.

generates the modified image. The combination of (prompt, input image, output image) forms the text-and-image-to-image data used to align multimodal models with proprietary models' image editing capabilities.

4 Janus-4o: Fine-Tuning with ShareGPT-4o-Image

To evaluate the effectiveness of ShareGPT-4o-Image, we fine-tune Janus-Pro on ShareGPT-4o-Image to develop **Janus-4o**. Since Janus-Pro only supports text-to-image generation, we enhanced its inference model and designed a corresponding text-and-image-to-image training method, enabling it to handle image editing tasks (Sec. 4.2). The complete training methodology for Janus-4o is shown in Figure 4. We employed a joint training approach for text-to-image and text-and-image-to-image generation (Sec. 4.3).

4.1 Text-to-Image Fine-Tuning

Training For text-to-image training, we fine-tune Janus-Pro under the original text-to-image generation setup. A text prompt is tokenized into $S = (s_0, s_1, \ldots, s_M)$, while the target image I is flattened into pixel patches and mapped via a codebook into image tokens $X = (x_0, x_1, \ldots, x_N)$. Both S and X are embedded through their respective text and image embedding layers and then fed into the LLM. The objective is to autoregressively predict the image tokens, with the loss:

$$\mathcal{L} = -\sum_{i=1} \log P_\theta \left(x_i \mid x_{<i}, S\right)$$

During fine-tuning, 10% of S is randomly masked to padding tokens to encourage pixel-level dependency modeling, in line with GPT-4o-style modeling.

Inference Inference follows the same next-token prediction process as Janus-Pro. For each token, the logit is computed as:

$$l_g = l_u + s(l_c - l_u)$$

where l_c is the conditional logit (with prompt), l_u the unconditional logit (prompt masked), and s a scaling factor. We use $s = 5$ and a temperature of 1.0.

4.2 Text-and-Image-to-Image Fine-Tuning: Unlocking New Capabilities

Training Janus-Pro does not natively support text-and-image-to-image generation. This task requires a semantic understanding of the input image to enable pixel-level modifications, making it necessary to incorporate both the image's semantic embedding and its tokenized representation. Given an input image $\hat{I}$, an image encoder $\mathcal{E}$ produces a semantic embedding $\mathcal{E}(\hat{I})$, and a codebook yields image tokens $\hat{X} = (\hat{x}_0, \hat{x}_1, ..., \hat{x}_N)$. These are concatenated with the prompt tokens S to form the model input. The training loss for generating the target image I is:

$$\mathcal{L} = -\sum_{i=1} \log P_\theta \left(x_i \mid x_{<i}, \mathcal{E}(\hat{I}), \hat{X}, S\right)$$

To avoid overfitting to the input image, 50% of $\hat{X}$ is randomly masked during fine-tuning.

Inference At inference, the input image is encoded into $\mathcal{E}(\hat{I})$ and $\hat{X}$, which are embedded into the input for autoregressive token generation. For each token, the logit is computed as:

$$l'_c = \frac{l_c + s' \cdot l_o}{1 + s'}$$

Table 1: Evaluation of *text-to-image* generation ability on GenEval benchmark. ★ indicates the base model before fine-tuning, and its results were rigorously reproduced by us.

Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
***Text-to-Image* Generation Models**							
LlamaGen (Sun et al., 2024a)	0.71	0.34	0.21	0.58	0.07	0.04	0.32
LDM (Rombach et al., 2022)	0.92	0.29	0.23	0.70	0.02	0.05	0.37
SDv1.5 (Rombach et al., 2022)	0.97	0.38	0.35	0.76	0.04	0.06	0.43
PixArt-α (Chen et al., 2023)	0.98	0.50	0.44	0.80	0.08	0.07	0.48
SDv2.1 (Rombach et al., 2022)	0.98	0.51	0.44	0.85	0.07	0.17	0.50
DALL-E 2 (Ramesh et al., 2022)	0.94	0.66	0.49	0.77	0.10	0.19	0.52
Emu3-Gen (Wang et al., 2024)	0.98	0.71	0.34	0.81	0.17	0.21	0.54
SDXL (Podell et al., 2023)	0.98	0.74	0.39	0.85	0.15	0.23	0.55
DALL-E 3 (Betker et al., 2023)	0.96	0.87	0.47	0.83	0.43	0.45	0.67
SD3-Medium (Esser et al., 2024)	0.99	**0.94**	**0.72**	0.89	0.33	0.60	0.74
Janus-Pro-7B★(Chen et al., 2025c)	**1.00**	0.85	0.53	**0.90**	0.69	0.58	0.76
Our Multimodal LLM							
Janus-4o	**1.00**	0.92	0.58	0.88	**0.70**	**0.70**	**0.80**
(Janus-Pro + ShareGPT-4o-Image)	(+0.00)	(+0.07)	(+0.05)	(−0.02)	(+0.01)	(+0.12)	(+0.04)

$$l_g = l_u + s \cdot (l'_c - l_u)$$

Here, l_c is the conditional logit with full input, l_o is the conditional logit with $\hat{X}$ masked, l_u is the unconditional logit with all inputs masked, and $s = 5$, $s' = 5$. The parameter s' controls deviation from the input image: lower values preserve more of the original, while higher values allow more creative changes. The sampling temperature is set to 1.0.

4.3 Joint Fine-Tuning

Training Details We fine-tune Janus-Pro-7B using the ShareGPT-4o-Image dataset, which includes 45K *text-to-image* and 46K *text-and-image-to-image* samples. The two tasks are jointly trained via random sampling over 3 epochs. All parameters of Janus-Pro-7B are fully fine-tuned, using a learning rate of 5×10^{-6} and a batch size of 128. The resulting model, **Janus-4o**, supports both *text-to-image* and *text-and-image-to-image* generation tasks. **Training was completed in 6 hours on a single 8×A800 GPU machine.**

5 Experiments

5.1 Experimental Setup

Benchmarks To assess the *text-to-image* generation capabilities of Janus-4o, we adopted the Janus-Pro evaluation protocol, leveraging GenEval (Ghosh et al., 2024) and DPG-Bench (Hu et al., 2024) for a comprehensive analysis of compositionality and semantic alignment. In addition, we evaluated Janus-4o's ability to generate images conditioned on both visual and textual inputs using ImgEdit (Ye et al., 2025b), focusing on single-turn edits.

Human Evaluation To evaluate alignment with human preferences, we conducted a human study comparing Janus-4o against Janus-Pro-7B (Chen et al., 2025c) and UltraEdit (Zhao et al., 2024). The comparison involved 58 *Text-to-Image* and 35 *Text-and-Image-to-Image* examples. Model outputs were randomly ordered, and the evaluators selected the preferred output based on instruction fidelity and visual clarity, or marked a tie. Results were reported as preference ratios.

5.2 Results and Analysis

***Text-to-Image* Performance** We evaluate the performance of *text-to-image* generation on GenEval and DPG-Bench. As shown in Table 1, Janus-4o achieves a 4-point improvement over Janus-Pro, highlighting the benefit of ShareGPT-4o-Image. In Table 2, Janus-4o achieves a score of 85.71 on DPG-Bench, representing a 1.6-point improvement over Janus-Pro. These results show that high-quality distillation data can continue to improve an existing image generation model's visual composition and dense-instruction following, even when only 45K text-to-image samples are used for adaptation.

> ***Takeaway 1:*** *The distillation data from leading image generation models, like GPT-4o, can effectively enhance the model's text-to-image generation capability, even with just 45K samples.*

Table 2: Evaluation of *text-to-image* generation ability on DPG-Bench benchmark. ★ indicates the base model before fine-tuning, and its results were rigorously reproduced by us.

Method	Global	Entity	Attribute	Relation	Other	Overall↑
***Text-to-Image* Generation Models**						
SDv1.5 (Rombach et al., 2022)	74.63	74.23	75.39	73.49	67.81	63.18
PixArt-α (Chen et al., 2023)	74.97	79.32	78.60	82.57	76.96	71.11
Lumina-Next (Zhuo et al., 2024)	82.82	88.65	86.44	80.53	81.82	74.63
SDXL (Podell et al., 2023)	83.27	82.43	80.91	86.76	80.41	74.65
Playground v2.5 (Li et al., 2024a)	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-DiT (Li et al., 2024b)	84.59	80.59	88.01	74.36	86.41	78.87
PixArt-Σ (Chen et al., 2024b)	86.89	82.89	88.94	86.59	87.68	80.54
Emu3-Gen (Wang et al., 2024)	85.21	86.68	86.84	90.22	83.15	80.60
DALL-E 3 (Betker et al., 2023)	90.97	89.61	88.39	90.58	**89.83**	83.50
SD3-Medium (Esser et al., 2024)	87.90	**91.01**	88.83	80.70	88.68	84.08
Janus-Pro-7B★(Chen et al., 2025c)	80.70	89.75	**90.04**	89.85	88.55	84.12
Our Multimodal LLM						
Janus-4o	**92.59**	90.61	89.51	**91.77**	89.01	**85.71**
(Janus-Pro + ShareGPT-4o-Image)	(+11.89)	(+0.86)	(−0.53)	(+1.92)	(+0.46)	(+1.59)

***Text-and-Image-to-Image* Performance** We evaluate the performance of *text-and-image-to-image* generation on ImgEdit-Bench. As shown in Table 3, Janus-4o achieves a score of 3.26 and remains competitive with editing models trained on much larger datasets, with particularly clear gains in the *Motion Change* and *Style Transfer* categories. More importantly, Janus-Pro originally lacks native support for *text-and-image-to-image* generation, while Janus-4o acquires this capability through only 91K synthetic samples and lightweight adaptation. This underscores the critical role of high-quality distillation data and highlights the effectiveness of ShareGPT-4o-Image.

> ***Takeaway 2:*** *Multimodal models with visual understanding and generation capabilities, such as Janus-Pro, can acquire image editing behavior through compact distillation data even if they were not originally designed for editing.*

Human Evaluation To provide a more comprehensive assessment of the model's generation quality, we conducted a human evaluation to compare the outputs. Specifically, we recruited 3 volunteers (not contributors to this paper) and asked each to create 20 text-to-image and 12 text-and-image-to-image prompts (with the option to search online for inspiration). After filtering, we retained 58 text-to-image and 35 text-and-image-to-image prompts. Volunteers then performed pairwise preference comparisons for the model outputs. They could choose the model they felt produced higher quality images or mark the outputs as a tie. The model names were kept anonymous, and the order of the comparisons was randomized. The final results were based on the majority vote.

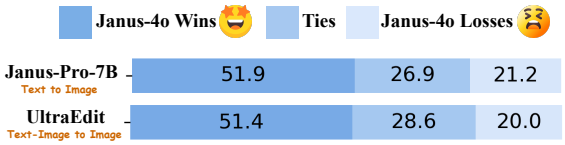


Figure 5: Human evaluation of Janus-4o's performance on *text-to-image* and *text-and-image-to-image* generation tasks. *Wins* indicate that the evaluator found Janus-4o superior in both instruction-following and image quality. *Ties* mean both models performed equally well, while *Losses* indicate that Janus-4o performed worse in those cases. Results were reported as preference ratios.

Human Evaluation Results The results in Figure 5 indicate that content generated by Janus-4o aligns more closely with human preferences than that of previous models. In the *Text-to-Image* task, Janus-4o produces images with noticeably higher visual quality compared to Janus-Pro. Furthermore, in the *Text-and-Image-to-Image* task, it exhibits a significantly improved ability to follow instructions, leading to more accurate image modifications based on the given prompts.

Table 3: Evaluation of *text-and-image-to-image* generation ability on ImgEdit-Bench. ★ indicates the base model before fine-tuning. **#Data** denotes training data size (K = thousand). Column abbreviations: Add. = Addition, Rmv. = Removal, Repl. = Replacement, Mot. = Motion Change, Style = Style Transfer, Bkg. = Background Change, Obj. = Object Extraction, Hyb. = Hybrid Edit, Avg. = average score.

Model	#Data↓	Add.	Rmv.	Repl.	Mot.	Style	Bkg.	Obj.	Hyb.	Avg.↑
***Text-and-Image-to-Image* Generation Models**										
Instruct-Pix2Pix (Brooks et al., 2022)	500K	2.29	1.49	1.93	1.51	3.54	1.67	1.33	1.48	1.91
AnySD (Yu et al., 2024)	2500K	3.12	2.34	2.71	3.31	3.27	2.37	1.82	2.07	2.62
UltraEdit (Zhao et al., 2024)	4100K	3.63	1.71	3.13	3.57	3.69	3.31	2.02	2.33	2.92
Step1X-Edit (Liu et al., 2025)	1000K	**3.90**	**2.61**	**3.45**	3.43	4.44	3.19	1.87	2.52	3.17
ImgEdit-E1 (Ye et al., 2025b)	1200K	3.82	2.40	2.80	3.21	4.38	**3.38**	**2.55**	**2.87**	3.17
Janus-Pro-7B★(Chen et al., 2025c)	0	*Not available*								-
Our Multimodal LLM										
Janus-4o (Janus-Pro + ShareGPT-4o-Image)	**91K**	3.60	2.28	3.27	**4.13**	**4.47**	3.32	2.28	2.74	**3.26**

Takeaway 3: *The improvement in image generation quality from distillation data is evident, and the gap in human evaluation is even more noticeable compared to the improvement in the benchmark datasets.*

Table 4: Comparison results between synthetic and real images. The experiment uses data from the image-first section of the text-to-image dataset, comparing real source images with synthetic images under the same prompt by training models.

	GenEval	DPG-Bench
Janus-Pro-7B	0.76	84.12
w/ synthetic images	0.77(↑ 0.01)	85.54(↑ 1.42)
w/ real images	0.71(↓ 0.05)	81.68(↓ 2.44)

Synthetic Images vs. Real Images To assess the importance of the distillation dataset, we compared images generated by GPT-4o with real images by training separate models. To ensure fairness in the training process, we used data from the image-first section of the text-to-image dataset, as described in Section 3.1. This section contains 22K entries, where each entry includes a prompt generated from a real image, resulting in a dataset with a prompt, a real image, and a GPT-4o-generated image. We split the data into two sets, training one model with real images and the other with synthetic images, both based on Janus-Pro-7B.

Ablation Experiment Results We conducted two comparison experiments: one trained on real images and the other on synthetic images (using the same prompts). The results are shown in Table 4. In this controlled setting, synthetic images improve the model's performance, while real images degrade both benchmarks. The higher quality and content consistency of the distillation dataset images (produced from the same model distribution) contribute to continued gains for a pre-trained model.

Takeaway 4: *In our controlled comparison, teacher-generated images provide more effective continued-training targets than the corresponding real images, likely due to stronger prompt-image consistency and higher visual quality.*

6 Conclusion

In this work, we introduce ShareGPT-4o-Image, a large-scale dataset that captures GPT-4o's advanced image generation capabilities in both *text-to-image* and *text-and-image-to-image* generation. Using this dataset, we develop Janus-4o, a MLLM capable of generating high-quality images from text alone or from combined image and text inputs. Janus-4o achieves substantial improvements in *text-to-image* generation and delivers competitive results on *text-and-image-to-image* tasks, underscoring the high quality and utility of ShareGPT-4o-Image. Benefiting from the efficiency of MLLM-based autoregressive image generation, Janus-4o can be trained in just 6 hours on a single machine with 8×A800 GPUs, achieving notable performance gains.

Limitations

- **Teacher-Model Dependence** ShareGPT-4o-Image benefits from the strong instruction-following and visual quality of GPT-4o-Image, which enables effective distillation with only 91K samples. However, because the teacher model is proprietary, its internal training data, filtering rules, and generation behavior are not fully transparent. The dataset may therefore inherit hidden biases or stylistic preferences from the teacher.

- **Coverage of Long-Tail Cases** Our data construction covers both *text-to-image* and *text-and-image-to-image* generation and spans diverse prompts, styles, and editing categories. Still, it cannot exhaustively cover all real-world requests, especially rare object combinations, culturally specific concepts, precise text rendering, complex spatial relations, and safety-sensitive scenarios.

- **Evaluation Scope** We evaluate Janus-4o on standard benchmarks and human preference comparisons, showing consistent gains over the base model. Nevertheless, these evaluations may not fully capture multi-turn editing, adversarial prompts, factual visual grounding, demographic fairness, or professional creative workflows. Broader evaluation with more diverse annotators would further strengthen the analysis.

- **Model Constraints** Janus-4o shows that an autoregressive MLLM can efficiently acquire both generation and editing capabilities. However, it still inherits limitations from Janus-Pro and discrete image-token generation, including possible artifacts, loss of fine details, weak small-text rendering, and imperfect preservation of input-image identity in challenging edits.

Despite these limitations, our results suggest that compact, high-quality synthetic supervision is a promising direction for data-efficient adaptation of open multimodal generators.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2022. Instructpix2pix: Learning to follow image editing instructions. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, and 1 others. 2025. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, and 1 others. 2025a. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024b. PixArt-Sigma: Weak-to-strong training of diffusion transformer for 4K text-to-image generation. *arXiv preprint arXiv:2403.04692*.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and 1 others. 2023. Pixart-$alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and 1 others. 2024c.

Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7346–7370.

Sixiang Chen, Jinbin Bai, Zhuoran Zhao, Tian Ye, Qingyu Shi, Donghao Zhou, Wenhao Chai, Xin Lin, Jianzong Wu, Chao Tang, and 1 others. 2025b. An empirical study of gpt-4o image generation capabilities. *arXiv preprint arXiv:2504.05979*.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025c. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024d. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Gheorghe Comanici, Eric Bieber, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*. Nano Banana refers to the image generation feature of Gemini 2.5 Flash.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *Preprint*, arXiv:2403.03206.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.

Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Advances in Neural Information Processing Systems*.

Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. 2024a. Seed-data-edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007*.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2024b. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2024. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, and 1 others. 2020. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. 2024. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. 2024a. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, and 1 others. 2024b. Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, and 1 others. 2025.

Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2024. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455.

Open AI. 2025. Introducing 4o image generation. https://openai.com/index/introducing-4o-image-generation/.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. 2025. Pico-banana-400k: A large-scale dataset for text-guided image editing. *arXiv preprint arXiv:2510.19808*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Team Seedream. 2025. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. 2024a. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

Yutao Sun, Hangbo Bao, Wenhui Wang, Zhiliang Peng, Li Dong, Shaohan Huang, Jianyong Wang, and Furu Wei. 2024b. Multimodal latent language modeling with next-token diffusion. *arXiv preprint arXiv:2412.08635*.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*.

Alex Jinpeng Wang, Dongxing Mao, Jiawei Zhang, Weiming Han, Zhuobai Dong, Linjie Li, Yiqi Lin, Zhengyuan Yang, Libo Qin, Fuwei Zhang, and 1 others. 2025a. Textatlas5m: A large-scale dataset for dense text image generation. *arXiv preprint arXiv:2502.07870*.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, and 1 others. 2024. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Yuhan Wang, Siwei Yang, Bingchen Zhao, Letian Zhang, Qing Liu, Yuyin Zhou, and Cihang Xie. 2025b. Gpt-image-edit-1.5 m: A million-scale, gpt-generated image dataset. *arXiv preprint arXiv:2507.21033*.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and 1 others. 2025. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, and 1 others. 2025a. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. 2025b. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, and 1 others. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. 2024. Anyedit: Mastering unified high-quality image editing for any idea. *arXiv preprint arXiv:2411.15738*.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, and 1 others. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*.

Hong Zhang, Zhongjie Duan, Xingjun Wang, Yuze Zhao, Weiyi Lu, Zhipeng Di, Yixuan Xu, Yingda Chen, and Yu Zhang. 2025. Nexus-gen: A unified model for image understanding, generation, and editing. *arXiv preprint arXiv:2504.21356*.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449.

Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. 2024. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, and 1 others. 2024. Lumina-Next: Making Lumina-T2X stronger and faster with Next-DiT. *arXiv preprint arXiv:2406.18583*.

A Potential Risks

ShareGPT-4o-Image and Janus-4o are released to support academic research on open multimodal generation. While the dataset improves access to high-quality distillation data, image generation and editing systems still require careful use. We identify the following potential risks.

- **Bias in Synthetic Images** Our prompt design aims to create diverse and high-quality data, but generated images may still reflect biases from the proprietary teacher model. Users should audit representation across demographic, cultural, and occupational attributes before applying the dataset or derived models in sensitive settings.

- **Misuse of Image Generation and Editing** Janus-4o enables useful text-to-image and text-and-image-to-image research, but similar capabilities could be misused to create misleading synthetic media, impersonation, nonconsensual edits, or rights-infringing content. We strongly discourage using the dataset or model for deception, harassment, surveillance, or identity manipulation.

- **Need for Safeguards** The dataset and model are intended for research use. Any downstream deployment should include content moderation, provenance disclosure or watermarking, human review for public-facing outputs, and compliance with applicable laws, consent requirements, and intellectual-property restrictions.

B Detailed Related Work

B.1 Instruction-guided Image Generation Models

Instruction-guided generation has recently gained momentum, with its main approaches centered on diffusion models (Song and Ermon, 2019; Ho et al., 2020; Rombach et al., 2022; Ho and Salimans, 2022; Ge et al., 2024b; Zhang et al., 2025; Chen et al., 2025a; Cai et al., 2025) and autoregressive generation models (Ramesh et al., 2021; Esser et al., 2021; Wu et al., 2025; Team, 2024; Zhou et al., 2024; Xie et al., 2024; Wang et al., 2024). Diffusion models (Ramesh et al., 2021; Yu et al., 2022) iteratively denoise random noise to produce high-fidelity, diverse images, capturing fine details and complex structures. Autoregressive models (Lu et al., 2024; Sun et al., 2023; Zhan et al., 2024; Sun et al., 2024a) discretize images into token sequences via vector quantization, enabling efficient, controllable generation and seamless integration with LLMs. To broaden the scope of applications, several works (Liu et al., 2025; Yu et al., 2024; Ye et al., 2025b; Zhao et al., 2024; Zhang et al., 2023; Brooks et al., 2022) investigate instruction-driven image editing, allowing models to create new images by modifying inputs according to specified instructions. *To the best of our knowledge, our model, Janus-4o, is the first autoregressive model capable of both image generation and image editing.*

B.2 Datasets for Image Generation

Early image-text pair datasets, such as MS-COCO (Lin et al., 2014) and TextCaps (Sidorov et al., 2020), provided natural language captions aligned with images and have been widely utilized in *text-to-image* generation tasks. To further enhance the performance and generalization capabilities of generative models, subsequent works (Changpinyo et al., 2021; Schuhmann et al., 2022; Wang et al., 2025a) collected larger-scale image-text pairs from the internet, enriching the accompanying textual descriptions. In addition to *text-to-image* synthesis, certain datasets (Brooks et al., 2022; Hui et al., 2024; Zhang et al., 2023; Ge et al., 2024a; Zhao et al., 2024; Yu et al., 2024; Ye et al., 2025b) focus specifically on image editing tasks, wherein models are required to modify images based on natural language commands. These tasks encompass operations such as action manipulation and style transfer, which support more fine-grained visual control and enable richer forms of multimodal interaction. However, the use of real-world images as training targets may introduce challenges, including low visual quality and inconsistency. Overall, there remains a significant lack of high-quality image generation datasets that can effectively distill the capabilities of the most advanced image generation models. *To the best of our knowledge, our work is the first to propose a distillation dataset with cutting-edge proprietary image generation models and demonstrates its effectiveness.* Subsequent works have followed our approach to construct distillation datasets (Qian et al., 2025; Ye et al., 2025a; Wang et al., 2025b).

B.3 Multimodal Large Language Models (MLLMs)

To enhance interactivity, Large Language Models (LLMs) have evolved beyond text, embracing a multimodal paradigm. With the emergence of GPT-4o (Achiam et al., 2023), researchers began enabling LLMs to understand images (Liu et al., 2023; Chen et al., 2024c; Bai et al., 2025; Team et al., 2023; Chen et al., 2024d) by aligning visual features with text using paired image-text data and pre-trained image encoders. Building on this, image generation became a natural extension. Early works like Vitron (Fei et al., 2024) treated it as tool use, where LLMs generated prompts for external models. Later, some approaches (Sun et al., 2024b; Wu et al., 2024; Zhan et al., 2024) enabled external decoders to generate images directly from LLM representations. Now, models like Janus support fully autoregressive image generation via image tokens (Henighan et al., 2020; Tian et al., 2024; Wu et al., 2025). These models offer unique advantages: (1) a simple end-to-end design with no extra components, and (2) native LLM understanding applied to image generation, enabling better instruction following. Janus-4o will be the first of its kind to support both *text-to-image* and *text-and-image-to-image* generation.

C Image Generation Categories

C.1 Categories

C.1.1 *Text-to-Image* Dimensions

Dimension	Sub-dimension
Background categories	a. abstract and surreal landscapes b. urban and architectural settings c. interior spaces
Style categories	a. cultural and historical aesthetics b. artistic movements and periods c. digital and graphic art styles
Lighting options	a. color temperature b. light direction c. light intensity and quality
Camera viewpoints	a. vertical perspective b. horizontal perspective c. shot distance and framing
Composition techniques	a. basic division and placement b. balance and symmetry c. framing depth and layers

Table 5: Text-to-Image Dimensions and Sub-dimensions.

Detailed Text-to-Image dimensions are shown in Table 5 (apart from **Object** dimension, whose sub-dimensions can be found in ImageNet (Deng et al., 2009)). The distribution among different categories can be found in Figure 6.

C.1.2 *Text-and-Image-to-Image* Categories

Category	Subcategory
I. Image editing and manipulation	a. Inpainting and replacement b. Element manipulation c. Background modification d. Attribute and effect manipulation
II. Style transfer and artistic transformation	a. Specific artist and art styles b. Medium and technique styles c. Genre theme and era shifting
III. Content augmentation and extension	a. Resolution detail and quality enhancement b. Image outpainting and inpainting for extension
IV. Structured generation and conditional control	a. From sketch lineart and edges b. From pose depth and segmentation
V. Creative ideation and iteration	a. Storyboarding and sequential generation b. Concept variation and exploration c. Hybridization and mashups

Table 6: Categories and subcategories for *Text-and-Image-to-Image* tasks.

Detailed categories and sub-categories for Text-and-Image-to-Image tasks can be found in Table 6. The distribution of the categories can be found in Figure 7.

C.2 Distribution of k

In this appendix, we explain the distribution used for sampling the number of object categories, k. The selection is based on a random sample ranging from 1 to 100, with weights applied to ensure diverse sampling patterns, especially when multiple objects interact.

C.2.1 Exponential Decay Distribution

We employ an exponential decay distribution to model the probability of selecting a higher number of categories. This approach ensures that the probability decreases exponentially as the number increases, making it suitable for scenarios where fewer categories are more common.

The distribution is defined as follows:

$$w(x) = \exp(-\lambda(x-1))$$

Where λ is the parameter controlling the decay rate; a larger λ results in faster decay.

The following figure illustrates the characteristics of this distribution:

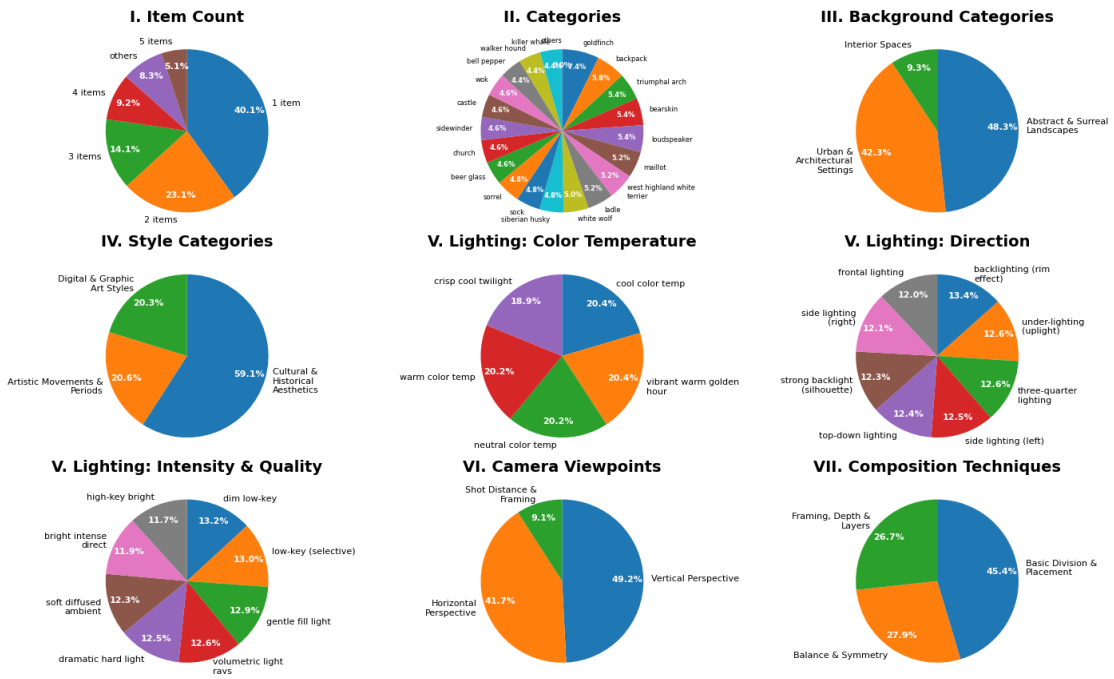


Figure 6: Text-to-Image Categories distributions.

Figure 7: Distribution of *Text-and-Image-to-Image* categories.

D Prompts for Generation

D.1 Image-First Prompt

Image-First Prompt

Please describe the main content of the image in one sentence. This sentence will be used as a prompt to regenerate the image, so it should clearly capture the key visual information. Only provide the sentence,no extra text.

D.2 Content Generation Process

Prompts for Content Generation

You are an expert document concept designer and text-to-image prompt generator. Your goal is to create concepts that are clear, professional, and highly appropriate for the intended use.

Your task is to perform the following steps IN ORDER:

STEP 1: Use the Given Theme and Choose a Core Document Category
- Use the following theme: "{selected_theme}"
- Choose ONE core document category that best fits this theme from the following list:
{CORE_DOCUMENT_CATEGORIES}
- Based on the chosen core category and theme, conceive and clearly state a MORE SPECIFIC document type or application.
For example:
[previous examples remain the same...]

STEP 2: Select an Appropriate Font Style and Visual Characteristics
- Choose and incorporate this font style: "{se-

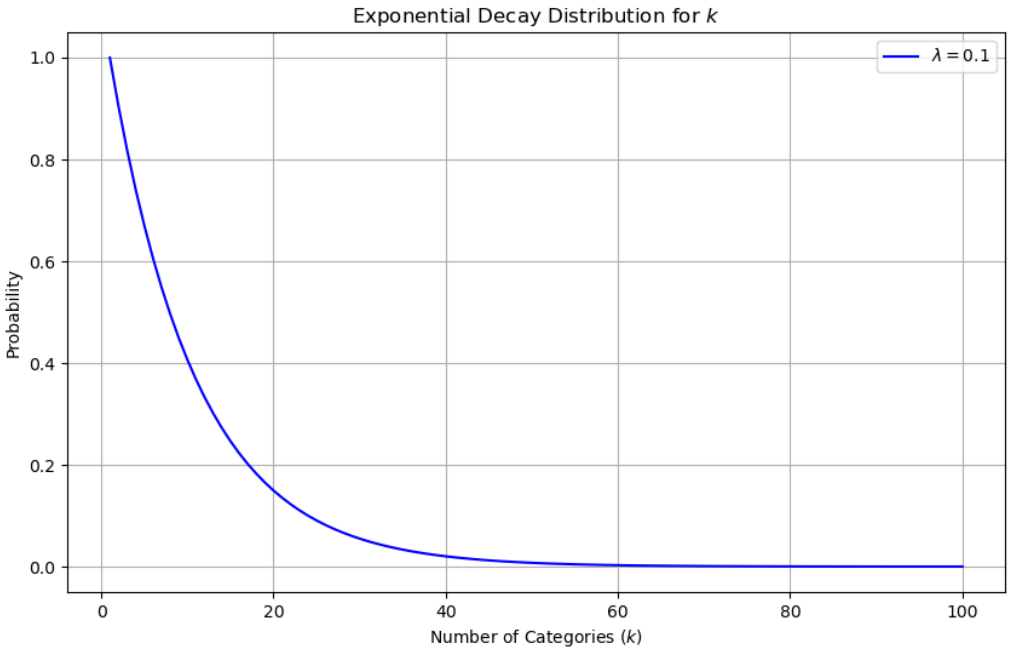


Figure 8: Illustration of the exponential decay distribution used for sampling k.

lected_font}"
- Based on the **Specific Document Type** from STEP 1 and the Theme, internally determine the most FITTING and EFFECTIVE visual style characteristics.
- **Priority is on CLARITY, PROFESSIONALISM, and FUNCTIONALITY for the specific document type.**
- Avoid overly ornate or unnecessarily complex styles unless the theme and specific type genuinely call for it (e.g., a historical document recreation).
- **Minimize generic "parchment paper" or "old paper" styles for most modern professional contexts.**
- The visual style should serve the content and the document's purpose.

STEP 3: Generate a Clear and Fitting Visual Template Prompt
- Now, generate a single, compelling text-to-image prompt (in English) for creating a visual template.
- This prompt MUST be based on the Theme AND the **Specific Document Type** (from STEP 1).
- The prompt MUST reflect a visual style that is **highly appropriate, clear, and professional** for the specific document type and theme (informed by STEP 2).
- The prompt should describe:
- Overall aesthetic (e.g., "clean and modern," "formal and academic," "bold and clear," "data-focused and organized").
- Layout and composition (e.g., "standard slide layout with title and content areas," "two-column academic paper layout," "grid-based infographic structure").
- Color palette (e.g., "corporate blues and grays," "high-contrast for readability," "theme-appropriate accent colors").
- Textures (if any: "smooth untextured background," "subtle professional texture like fine linen," "no distracting textures").
- Placeholder elements and their style, suitable for the specific document type (e.g., "placeholders for a large title, speaker name, and event logo for a title slide").
- Do NOT include any specific readable text in this visual template prompt itself.
- The prompt should be a single, coherent sentence.

STEP 4: State Your Outputs Clearly
- Provide your outputs in the following exact format, with each item on a new line:
Visual Template Prompt: [The visual template prompt you generated in STEP 3]
Conceived Theme: [The theme you conceived in STEP 1]
Document Type: [The CORE DOCUMENT CATEGORY chosen in STEP 1, followed by the SPECIFIC document type/application you conceived in STEP 1, e.g., "Slide - Title Slide for a Tech Conference"]

Now, follow these steps carefully, focusing on CLARITY, PROFESSIONALISM, and APPROPRIATENESS for the document type, and generate a new set:

D.3 Meta Prompts for Image Text Instruction

Meta Prompts for Image Text Instruction

You are an AI assistant that creates a single, concise, complete, and descriptive English sentence to be used as a prompt for image generation models.
This sentence should beautifully capture the essence, mood, and key visual elements of the provided input image.
If a 'Base Task' is provided, consider integrating it naturally into this single descriptive sentence.
However, prioritizing the image's core visual appeal and

the formation of a well-structured, impactful sentence is more important than rigidly following the task.
"If the task doesn't fit well or makes the sentence awkward or verbose, focus on crafting the best single descriptive sentence for the image itself.
Output ONLY this single English sentence. No explanations, no preambles, no bullet points, no numbered lists. Just one sentence.
For example, for an image of a misty forest (task: 'add creature'): 'A mystical deer steps softly through the sun-dappled, misty ancient forest.'
Or if the task is awkward: 'Sunlight filters mysteriously through the tall trees in the quiet, misty forest.'
For a portrait (task: 'change background to space'): 'Her serene gaze is complemented by a backdrop of swirling cosmic stardust.'
Or without task: 'A soft light illuminates her thoughtful expression in this intimate portrait.'

E Document Pipeline

E.1 Pipeline

Document Prompts Generation. To generate realistic document-style images (e.g., slides, posters, reports), we implemented a specialized workflow:

1. **Content Generation**: The LLM first generates plausible textual and structural content for the document (e.g., titles, paragraphs, data snippets). This is guided by meta-prompts tailored to different document types (see Appendix E.2 for examples).

2. **Prompt Synthesis**: Based on this generated content, the LLM then formulates a detailed textual prompt describing the desired visual layout, typography, color scheme, and overall appearance of the document image. The meta-prompts used for this stage are also provided in Appendix D.2.

This specialized approach helps the prompt for document images to capture the unique compositional and semantic nuances inherent to this image category.

E.2 Core Document Categories

- Slide
- Paper
- Document
- Infographic/Chart
- Poster