DreamOmni: Unified Image Generation and Editing

Bin Xia [1], Yuechen Zhang [1], Jingyao Li[1], Chengyao Wang [1],
Yitong Wang[2], Xinglong Wu[2], Bei Yu[1], and Jiaya Jia [3]
[1] CUHK, [2] ByteDance Inc, [3] HKUST
https://zj-binxia.github.io/DreamOmni-ProjectPage/
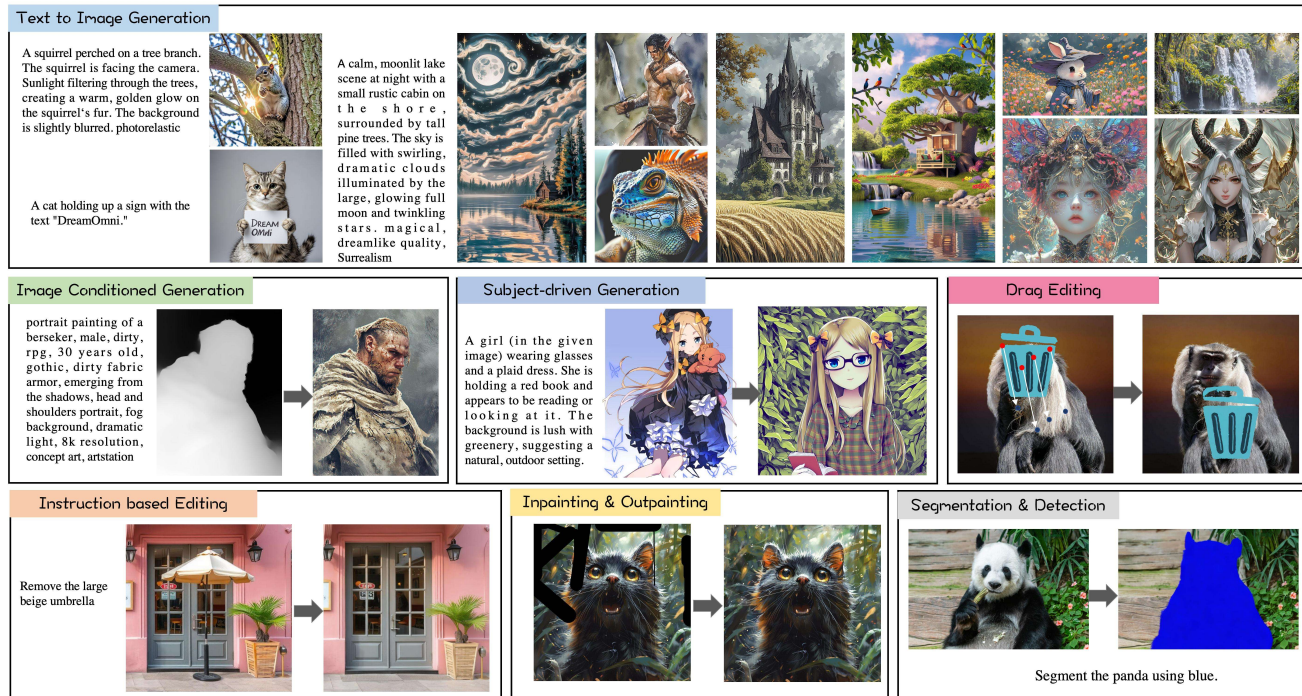
Figure 1. The gallery of DreamOmni. DreamOmni, as a native unified image generation and editing model, can handle various tasks.

Abstract

projectpagepCurrently, the success of large language models (LLMs) illustrates that a unified multitasking approach can significantly enhance model usability, streamline deployment, and foster synergistic benefits across different tasks. However, in computer vision, while text-to-image (T2I) models have significantly improved generation quality through scaling up, their framework design did not initially consider how to unify with downstream tasks, such as various types of editing. To address this, we introduce DreamOmni, a unified model for image generation and editing. We begin by analyzing existing frameworks and the requirements of downstream tasks, proposing a unified framework that integrates both T2I models and various editing tasks. Furthermore, another key challenge is the efficient creation of high-quality editing data, particularly for instruction-based and drag-based editing. To this end, we develop a synthetic data pipeline using sticker-like elements to synthesize accurate, high-quality datasets efficiently, which enables editing data scaling up for unified model training. For training, DreamOmni jointly trains T2I generation and downstream tasks. T2I training enhances the model's understanding of specific concepts and improves generation quality, while editing training helps the model grasp the nuances of the editing task. This collaboration significantly boosts editing performance. Extensive experiments confirm the effectiveness of DreamOmni. The code and model will be released.

1. Introudction

Recently, text-to-image (T2I) generative foundation models [6, 15, 27, 39, 44, 47] have made remarkable progress, driving the development of various downstream applications such as image editing, video generation, and more.

Currently, T2I foundation models face two significant challenges: **(1)** Adapting these models for downstream applications often requires the integration of various plug-ins (such as ControlNet [62] and IP-adapter [60]) in different ways, or the extension of input channels (*e.g.*, SD-inpainting [44], InstructP2P [8]). This reliance on specialized frameworks hampers multi-task generalization and complicates deployment. **(2)** High-quality and accurate editing data is difficult to obtain, including instruction-based editing, drag editing, and subject-driven generation data. In this paper, we propose to unify T2I models with multiple editing tasks, such as instruction-based editing, inpainting & outpainting, drag editing, and reference image generation within a single framework. Additionally, we introduce an efficient synthetic data pipeline for efficiently and accurately constructing editing data, facilitating the training of native unified generation and editing model.

We began by exploring fundamental model frameworks to develop a unified framework that is both highly effective and efficient, achieving rapid training convergence. **(1)** We aligned certain frameworks, such as Unet [44] and DIT [38], to have similar parameter sizes and runtime settings for fair comparison (Fig. 3). Our analysis showed that DIT's effectiveness comes from concentrating most of its computational and parameter load on latent at $2\times$ downsampling size, whereas Unet distributes more of its operations at smaller scales. This allocation makes DIT comparatively superior, as concentrating computations at the $2\times$ downsampled latent achieves a better trade-off. However, we observed that Unet's residual connections significantly improve the model's training convergence speed. Inspired by the efficiency of high-resolution latent computations, we also incorporated additional residual convolutional networks for the input-scale latent. **(2)** We replaced the original text encoder with a vision-language model (VLM) to unify visual-language prompt encoding. This encoded prompt was then concatenated with the noisy latent, allowing for integrated computation within the DIT framework.

Next, we need to obtain accurate training data efficiently for our unified model. For various types of editing, such as instruction-based editing, it's challenging to create and filter pairs of data that precisely match the instructions while avoiding detail corruption. We have discovered that the key to effective editing lies in helping the model understand the meaning of the editing operations, rather than learning specific concepts (a capability the model already possesses in T2I training). Therefore, we propose a synthetic collage data pipeline that easily scales up to generate accurate and diverse edited data, enabling models to better understand the nuances of specific editing tasks (Fig. 2 (b)). **(1)** For instruction-based editing, we constructed synthetic data for three main categories: addition, removal, and replacement. **(2)** For drag editing, we utilized synthetic data to cre-

ate scaling, translating, and rotating editing data. **(3)** For inpainting and outpainting, we randomly generate masks for each given image. **(4)** For reference image generation, We divided the tasks into subject-driven generation and image-conditioned generation (*i.e.*, ControlNet-like generation) tasks. For subject-driven generation, we used a canvas filled with stickers, allowing the model to reference a specific sticker for generation, which helps train the model to extract the demanded objects or details from given images for generation. For the image-conditioned generation, we created canny, depth, and segment maps, similar to ControlNet. **(5)** We use synthetic data to enhance the model's T2I accuracy in responding to attributes such as quantity, position, relationships, color, shape, and text. Concretely, we randomly arrange different quantities of stickers, text, and geometric shapes on a canvas, and obtain precise descriptions of their quantity, position, relationships, and colors based on their exact coordinates.

During training, thanks to our unified framework that integrates T2I with multiple editing tasks, we can easily train T2I data alongside various types of editing data, obtaining a native unified image generation and editing model, called DreamOmni. The editing tasks benefit from the T2I data, which helps prevent the model from forgetting specific concepts and generation quality decline, while the T2I task can leverage synthetic data to enhance instruction following.

- We conducted an analysis of existing model frameworks under fair settings and, drawing on the characteristics of different tasks, proposed an efficient and powerful unified image generation and editing framework, DreamOmni.
- We introduced a synthetic collage data pipeline to tackle the current inefficiency and difficulty in creating and filtering high-quality editing data. Furthermore, we utilized the synthetic collage data pipeline to enhance the accuracy of the T2I model's output. Experimental results demonstrate that synthetic data is an effective, high-quality, and cost-efficient method to scale up data for achieving unified image generation and editing training.
- After unified training with T2I data and various synthetic datasets, our DreamOmni showcases competitive performance on T2I generation and various editing tasks.

2. Related Work

Text-to-image Diffusion Model. Diffusion models [1, 5, 14, 21, 23, 24, 35, 52, 53] have emerged as highly effective methods for image generation, outperforming previous generative models like GANs [20] and VAEs [26]. The latent diffusion model [44], also known as Stable Diffusion, enhances the diffusion process within the image latent space and utilizes cross-attention to guide image generation through text input. This approach enhances the usability of the image generation model, driving its widespread adoption. Subsequently, more advanced text-to-

(a) The framework of DreamOmni
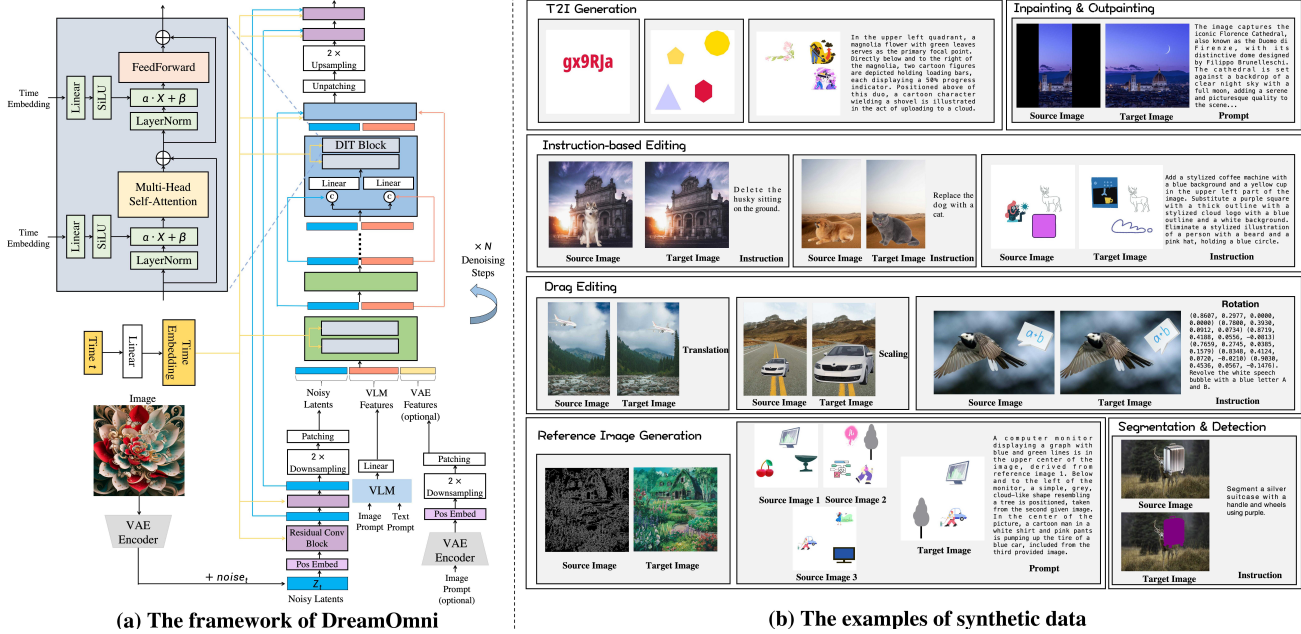
(b) The examples of synthetic data

Figure 2. The overview of DreamOmni. **(a)** The DreamOmni framework supports unified image generation and editing, with fast training convergence and powerful performance. **(b)** To overcome the difficulty and inefficiency in data creation and filtering for image editing, we propose a collage-based synthetic data pipeline. This pipeline enables the efficient creation of data for various editing tasks, such as adding, deleting, and replacement operations in instruction-based editing, as well as translation, scaling, and rotation in drag editing. Additionally, it supports reference image generation and segmentation & detection. Furthermore, our synthetic data pipeline enhances the accuracy of T2I generation. Due to space limitations, we have optionally shown the corresponding prompts or instructions for these cases.

image (T2I) models were developed, such as Imagen [47], DALLE [43], and others [3, 6, 13, 16, 39]. Recently, the DIT frameworks [4, 38] has become increasingly popular as more models [10, 15, 31] transition from the Unet framework [45], thanks to its straightforward structure and outstanding compatibility with multimodal inputs. However, the current T2I model fails to account for the design of downstream editing tasks, necessitating numerous specialized adaptations. For instance, Instructp2p [8] requires an expansion of input channels, while ControlNet [62] and IP-adapter [60] rely on additional plugins and various methods for injecting reference image information. These fragmented designs are inconsistent with the trend toward unified models, complicating deployment and impeding the joint training of T2I and editing tasks. In this paper, we will analyze the existing T2I frameworks and propose a more efficient and powerful unified generation and editing model.

Multimodal-guided Image Editing and Generation offers richer guiding information and has broader applications [2, 7, 9, 12, 37, 54, 55] than T2I models. It encompasses instruction-based editing [8, 18, 25, 49, 61], inpainting & outpainting [34, 57, 59], drag editing [32, 51], and reference image generation [28, 29, 40, 62, 63]. A significant challenge in editing, especially in instruction-based and drag editing, lies in creating and filtering accurately paired editing data. For example, InstructP2P [8] utilizes finetuned GPT-3 and a Prompt-to-Prompt strategy [22] to

generate instruction-based editing datasets, but it achieves a success rate of under 15% with frequent artifacts and detail loss. Additionally, filtering high-quality editing data is a persistent challenge, complicating scaling up. Magicbrush [61] addresses this by employing people to create and filter editing data, but this manual method limits data volume. Furthermore, for drag editing, most methods [32, 51] are training-free, which restricts their performance and efficiency. To this end, InstaDrag [50] leverages video data to construct drag editing data. However, InstaDrag still faces challenges with inefficiency in filtering valid data from video. Moreover, for subject-driven generation [28, 36], DreamBooth [46] requires training on several subject-related images before each inference, which is inconvenient and inefficient. Some works [30, 58] collect multiple photos of the same person from the internet to build datasets for training more efficient and powerful subject-driven generation models. However, the process of gathering and filtering such data still presents a challenge, limiting the scaling up of the dataset. In this paper, we propose a simple, efficient, and accurate synthetic collage data pipeline to create quite a few data for unified training.

3. Methodology

The unification of multiple tasks is a trend and pursuit in the field of computer vision and AI, which not only en-

hances model usability and reduces deployment complexity but also enables collaborative training that fosters synergies between tasks. However, current T2I foundation models are primarily designed specifically for T2I and often overlook the potential for integration with other tasks, such as various image editing tasks. To this end, we propose DreamOmni, a unified model for image generation and editing. We design and train DreamOmni from three aspects. **(1)** In Sec. 3.1, we compare various frameworks in a fair setting and design a powerful and fast training converging framework that supports unified multi-tasking based on the characteristics of different tasks. **(2)** In Sec. 3.2, we aim to propose a convenient, efficient, and accurate synthetic data pipeline for data scaling up to facilitate multi-task unified training and enhance the model's instruction-following ability. **(3)** In Sec. 3.3, we introduce the training scheme for DreamOmni. Native unified training of T2I and various editing tasks prevents concept forgetting and generation quality decline, while enhancing the model's editing and prompt-following capabilities.

3.1. Framework

In this part, we aim to design a unified and powerful image generation and editing framework. Currently, different editing models often have distinct structure designs. For example, IP-adapter [60] and BLIP-Diffusion [28] inject information through cross-attention to maintain subjects. In contrast, InstructP2P [8] achieves editing consistency by adding different numbers of input channels for models. These structures are tailored for specific tasks and lack generalizability. To this end, as shown in Fig. 2 (a), we concatenate VLM features with noisy latent and input them into a DIT Block for joint multi-head self-attention operations. After that, the VLM features and noisy latent are processed by the FeedForward modules. This allows the model to autonomously learn any level features (from overall consistency to subject consistency) for editing and generation. Notably, for FeedForward modules, we separate the VLM features and noisy latent, passing them through two distinct FeedForward modules with the same network structure. Additionally, instead of using CLIP [41] or T5 [42] as the text encoder, we introduce a Vision-Language Model (VLM) that enables joint understanding and encoding of both image and text prompts. For tasks that require high consistency, such as instruction-based editing and drag editing, we input the VAE encoding of the source image into the DIT model to ensure strong consistency between the non-edited regions of the output image and the source image.

In the current framework design, some works, such as DIT [38], are compared under label-conditioned generation rather than T2I. However, T2I is inherently more complex than label-based generation, as it requires the integration and understanding of complex prompts. Additionally, many

T2I models [15, 44], such as SDXL [39], are trained with different model sizes, datasets, and training settings. This variability makes it challenging to assess the impact of different model components on overall performance. Moreover, SDXL incorporates many Transformer blocks within its UNet structure. So, why does DIT outperform SDXL? To address this, we conducted extensive experiments as shown in Fig. 3. We observe that DIT surpasses Unet because it allocates most of its computation to $2\times$ downsampled latent, whereas Unet allocates a greater proportion to $4\times$ downsampled latent. Since attention operations on $1\times$ latent cause memory burden, we further employ residual Conv blocks to refine generation details for $1\times$ latent.

Additionally, we observed that the use of long connections in the UNet framework can significantly accelerate the model's training convergence without compromising performance. As shown in Fig. 2 (a), we concatenate early and later features along the channel dimension and apply a linear layer to combine the two features. Notably, linear layers used for VLM features and noisy latent are different.

3.2. Synthetic Data

In addition to the unified framework, we also require quite a few data to support joint training. While T2I data is readily available, creating and filtering accurate, high-quality data for tasks like instruction-based editing is far more challenging. To address this, we introduce a synthetic collage data pipeline that efficiently and accurately generates the required editing data. As shown in Fig. 2 (b), our pipeline encompasses six tasks. Notably, this is not the full extent of our synthetic pipeline's capabilities; it is also capable of handling more complex task combinations.

- *T2I generation.* As shown in Fig. 2 (b), in addition to the conventional T2I data, we further enhance the model's performance on the T2I by incorporating synthetic data, specifically focusing on improving the text, shape, position, quantity, and color generation. Specifically, for the text, we randomly generate words or short phrases on a blank canvas using a variety of fonts, colors, thicknesses, and sizes. For the shape and quantity, we randomly create geometric shapes with different quantities, colors, and sizes, and arrange them on the canvas. Based on these attributes and their positions, we generate accurate prompts, which are then refined using a LLM. Furthermore, we use a diverse set of stickers and segmentation data for synthesis, placing them on the canvas and calculating their precise spatial relationships. These prompts are then created and further polished by the LLM to produce descriptions that are more natural.

- *Inpainting & Outpainting.* We randomly generate masks for smearing, blocks, and image edges. Notably, during training, in addition to feeding the masked image and its corresponding mask into the VLM for encoding, we also

include the image caption with a 50% probability.

- *Instruction-based editing.* We categorize the task into three operations: addition, removal, and replacement. For both removal and replacement, we randomly select a background image and an object image to create the source image. In the removal case, the target image is simply the background image. For replacement, the target image is generated by replacing the object in the same location with a different object. Notably, for the addition, due to the need for the added object to be placed in a contextually appropriate position relative to the background, we use a blank background in the paper.

- *Drag editing.* We categorize the data into three types: translation, scaling, and rotation. Notably, Instadrag [50] treats each pair of drag points as a separate image, which is sparse and impractical due to the fixed number of drag points required. Therefore, we represent each drag point using the format (x, y, dx, dy) as a prompt input, where x and y denote the coordinates of the drag points in the source image, and dx and dy represent the translation vector. Furthermore, we normalize these coordinates by dividing them by the image's width or height.

- *Reference image generation.* We categorize the data into two types: image-conditioned generation, akin to ControlNet [62], and subject-driven generation. For image-conditioned generation, we begin by selecting high-quality images and creating corresponding canny maps, depth maps, and segmentation masks as source images for training. For subject-driven generation, we synthesize source images and then randomly select objects within these images to create target images. The model is trained to generate new content based on specific attributes referenced from the source image, allowing for the flexible generation of varied scenes and subjects.

- *Segmentation & detection.* We randomly selected a background image and an object image. These images are then composited together to create a source image. Following this, we apply color manipulation or draw a bounding box around the object region based on the alpha channel of the object image, obtaining the target image.

Our synthetic collage data pipeline is efficient and precise, enabling the generation of billions of diverse images for large-scale pretraining and fine-tuning of DreamOmni.

3.3. Model Training

After careful consideration, we chose a $2.5B$ parameter size for DreamOmni's DIT model. This size strikes a good balance between being user-friendly and ensuring powerful performance. For the VLM encoder, we directly adopt the Qwen2-VL [56] $7B$ model, based on three reasons: **(1)** It supports image inputs of arbitrary resolution, **(2)** It delivers strong model performance, and **(3)** It is released under a permissive open-source license. The VLM features are derived

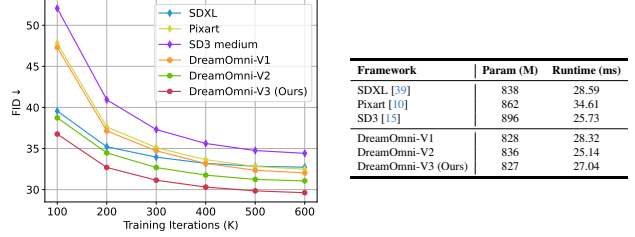| Framework	Param (M)	Runtime (ms)
SDXL [39]	838	28.59
Pixart [10]	862	34.61
SD3 [15]	896	25.73
DreamOmni-V1	828	28.32
DreamOmni-V2	836	25.14
DreamOmni-V3 (Ours)	827	27.04

Figure 3. Comparison of different frameworks. The left figure shows the FID comparison among different frameworks, while the right table shows their number of parameters and runtime.

Table 1. Comparisons on GenEval [19]. Our model outperforms all current open models (including SOTA SD3-Medium [15]).

		Objects					
Model	Overall	Single	Two	Counting	Colors	Position	Color Attribution
SD 1.5 [44]	0.43	0.97	0.38	0.35	0.76	0.04	0.06
PixArt-alpha [10]	0.48	0.98	0.5	0.44	0.8	0.08	0.07
SDv2.1	0.5	0.98	0.51	0.44	0.85	0.07	0.17
SDXL [39]	0.55	0.98	0.74	0.39	0.85	0.15	0.23
SD-cascade	0.52	0.98	0.57	0.47	0.87	0.08	0.15
SD3-Medium [15]	**0.70**	**0.99**	**0.84**	0.63	**0.88**	0.28	**0.55**
DreamOmni (Ours)	**0.70**	**0.99**	0.81	**0.65**	**0.88**	**0.34**	0.54

from the penultimate layer of Qwen2-VL. Additionally, we use the FLUX-schnell's VAE as DreamOmni's VAE, which retains more latent channels, enabling the model to capture finer image details. Furthermore, we optimize DreamOmni using Rectified Flow [33], which performs the forward process by linearly interpolating between noise and data along a straight trajectory. we train DreamOmni using loss $\mathcal{L}$:

$$\mathcal{L} = \mathbb{E}\left(\|(\mathbf{z} - \epsilon) - v_\theta\left(\mathbf{z}_t, \mathbf{c}, t\right)\|_2^2\right), \quad (1)$$

where $\mathbf{z}_t = t\mathbf{z} + (1 - t)\epsilon$ represents the noised feature map at timestep t. Ground truth images are encoded by VAE into latent space to derive $\mathbf{z}$. Here, $\epsilon \in \mathcal{N}(0, \mathbf{I})$ represents Gaussian noise, and v_θ refers to the DIT model. $\mathbf{c}$ indicates the conditional information.

The T2I training dataset consists of $125M$ images, including both the LAION dataset $(103M)$ [48] and our collected data $(22M)$. These images have been annotated using InternVL2 [11]. For synthetic data, such as T2I, instruction editing, inpainting & outpainting, drag editing, and reference image generation, we generated $12M$ images respectively, bringing the total to approximately $60M$ images. Additionally, for segmentation & detection, we created $8M$ images. Prior to training, to improve training speed usage, we first encoded prompts into VLM features.

Our training process is divided into three stages. In the first stage, we train on images with 256×256 size, a batch size of 2048, a learning rate of 1×10^{-4}, and $377K$ iterations. In the second stage, we train on images with 512×512 size, a batch size of 1024, a learning rate of 5×10^{-5}, and $189K$ iterations. In the final stage, we select 12M high-quality T2I data and randomly sample 1M high-quality images from each type of synthetic data for training.

| Prompt | SD1.5 | SD-Cascade | SDXL | SD3-Medium | DreamOmni (Ours) |

A heartwarming 3D rendered scene of an elderly farmer and a tiny orange kitten dressed in matching denim overalls. The farmer, with a gentle smile, walks alongside the kitten in a lush, green garden filled with thriving plants, showcasing a fruitful harvest. The intricate details of the overalls and the farmer's worn, weathered face tell a story of years spent tending to the land.
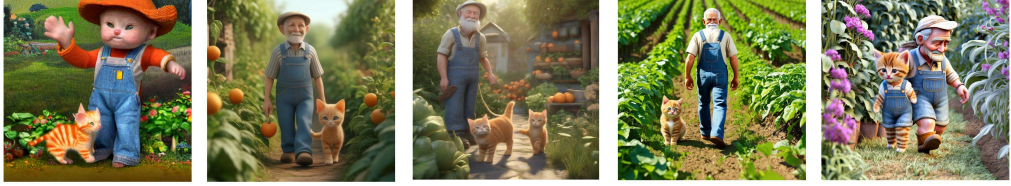
Deep in the enchanted forest lives a woman who is the moon fairy. Her long blonde hair shines in the starlight, tangled with her flowers that glow with a soft blue glow. Her eyes are the color of the night and shine with the magic of the night. The fairy wears a dress made of moon petals, woven with threads of moonlight that shine with an iridescent glow, a crown of stars adorns her head, shining with the light of the full moon that illuminates the forest. Her wings are translucent like glass, with a pale glow reminiscent of the glow of the moon. HD, 16K, photo, cinematic, poster
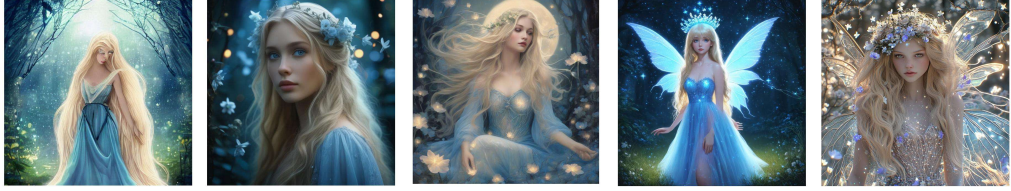
A unique, intricately detailed creature resembling a reptile, possibly a lizard or a gecko. It has a vibrant blue and green scaled body, with large, round, and expressive eyes that are a deep shade of blue. The creature is adorned with a floral crown made of various flowers, including pink roses and blue blossoms, as well as pinecones. The backdrop is a soft, blurred forest setting, suggesting a serene and mystical ambiance.

Figure 4. Visual comparison on **T2I generation**. Compared to other competitive methods (including SD3-Medium [15], SDXL [39], SD-Cascade, and SD1.5 [44]), our DreamOmni not only better adheres to user prompts but also generate more visually appealing results with delicate details, elegant composition, and so on.
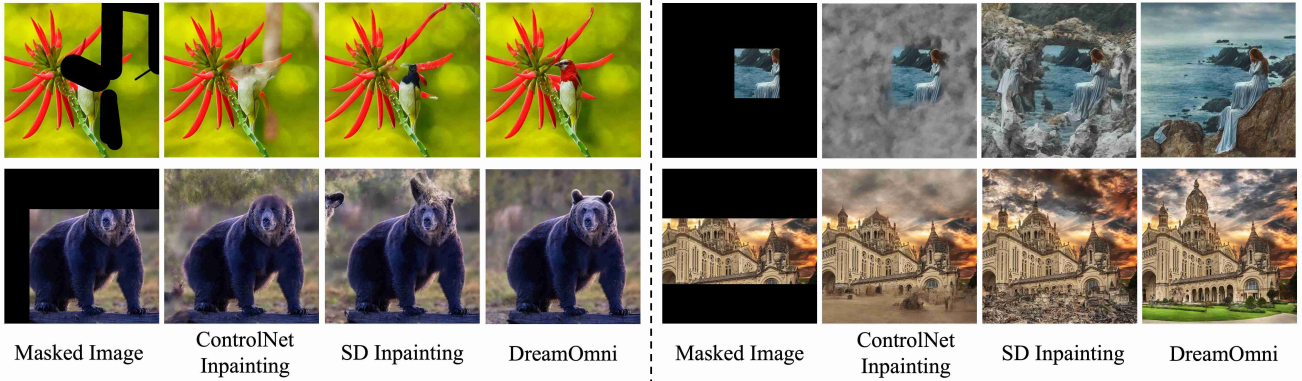

| Masked Image | ControlNet Inpainting | SD Inpainting | DreamOmni | Masked Image | ControlNet Inpainting | SD Inpainting | DreamOmni |

Figure 5. Visual comparison on **inpainting** & **outpainting** between DreamOmni, ControlNet-Inpainting [62] and SD-inpainting [44].

We train models with 1024×1024 size, a batch size of 256, a learning rate of 2×10^{-5}, and $140K$ iterations. All experiments are conducted on 64 A100 GPUs. Additionally, to enable the model to generate images at varying resolutions, similar to the approach used in SDXL [39], we divide the images into 31 buckets based on their aspect ratios, ranging from $4:1$ to $1:4$, during training.

4. Experiments

Evaluation on Framework. We compared several T2I model frameworks under similar settings to identify effective components. Using the same VAE, CLIP text encoder, parameters, runtime, and LAION training/testing datasets, we evaluated the Unet-based SDXL [39], DIT-based Pixart [10], SD3-Medium [15], and our developed DreamOmni variations. Notably, as shown in Fig. 3, to facilitate comparison, we do not use the full $2.5B$ parameters of DreamOmni, but instead adjust the parameters of all models to $0.85B$. DreamOmni-V1 has two downsampling layers ($2\times$ and $4\times$) like SDXL but lacks an Unet connection. DreamOmni-V2 builds on DreamOmni-V1 by adding the Unet connection, while DreamOmni-V3 takes DreamOmni-V2 further by focusing all DIT operations on $2\times$ downsampled latent. **(1)** Models with the Unet connection (SDXL, DreamOmni-V2, DreamOmni-V3) showed significantly faster convergence than those without (SD3-Medium, DreamOmni-V1). DreamOmni-V3, notably, converges four times faster than the SD3-Medium, enhancing
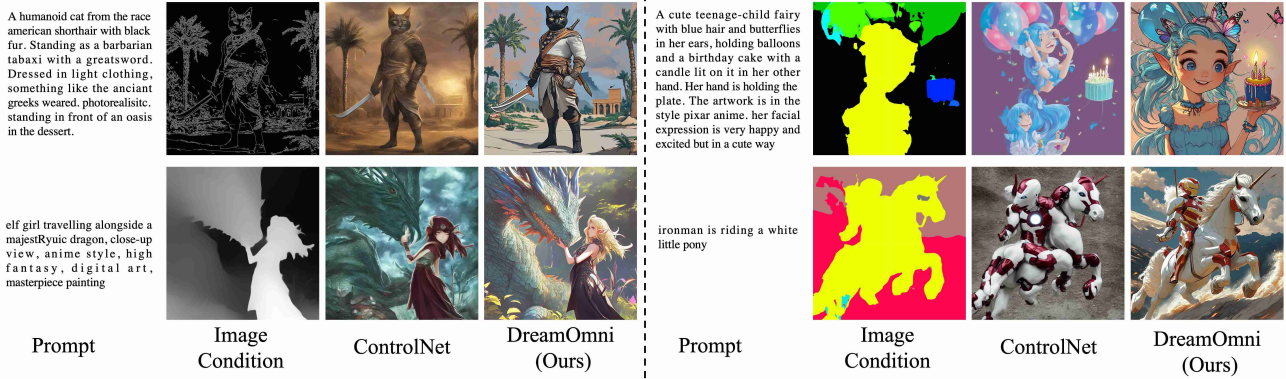
Figure 6. Visual comparison on **image-conditioned generation**. Our DreamOmni excels at using canny, depth, and segmentation maps as conditions for image generation. Compared to the classic ControlNet [62], DreamOmni not only follows the user's prompts and image conditions more accurately but also generates content and color schemes that are visually more pleasing.

Figure 7. Visual comparison on **subject-driven generation**. Compared with BLIP-Diffusion [28] and IP-adapter [60], DreamOmni excels at accurately following user prompts while preserving the specified subject. Furthermore, its success across both photo and anime cases underscores DreamOmni's generalization capabilities.

both training and fine-tuning efficiency. **(2)** Comparing DreamOmni-V3 and DreamOmni-V2, we found that concentrating DIT block computations on higher-resolution latent ($2\times$) is more cost-effective.

Evaluation on T2I generation. As shown in Tab. 1, our synthetic data significantly enhances DreamOmni's T2I generation capabilities in aspects such as quantity, color, and position, enabling our model to achieve SOTA results on GenEval [19]. Notably, SD3-Medium [15] is a $2B$ open-source SOTA T2I model, which has a similar parameters as our DreamOmni. Furthermore, qualitative results are shown in Fig. 4. We can see that DreamOmni's outputs are not only more visually pleasing but also align more accurately with the given prompts.

Evaluation on Inpainting. We compare our DreamOmni with ControlNet-Inpainting [62] and SD-Inpainting [44] on our high-quality evaluation datasets to evaluate its performance. The quantitative results, presented in Tab. 2, demonstrate that DreamOmni significantly outperforms both ControlNet-Inpainting and SD-Inpainting, highlighting its superior generation quality and coherence. Visual results (Fig. 5) further emphasize that DreamOmni ex-

cels in generating fine details, surpassing both ControlNet-Inpainting and SD-Inpainting. Additionally, DreamOmni effectively handles large mask holes, generating realistic content instead of blurry and incoherent outputs.

Evaluation on Reference Image Generation. For image-conditioned generation, we conducted a comparison with ControlNet [62] on canny, depth, and segmentation image conditions. Visual results are presented in Fig. 6. These results reveal that, across all tested conditions, our DreamOmni significantly outperforms ControlNet. Our method not only adheres more faithfully to both the image condition and the prompt but also demonstrates enhanced visual quality with better composition and richer detail.

For subject-driven image generation, we compared our method with competitive approaches, including BLIP-Diffusion [28] and IP-Adapter [60]. To demonstrate DreamOmni's powerful generalization, we validated their performance on both anime and photographic images. Visual results are shown in Fig. 7. Compared to other methods, our DreamOmni not only preserves the specified subject effectively but also follows the prompt well.

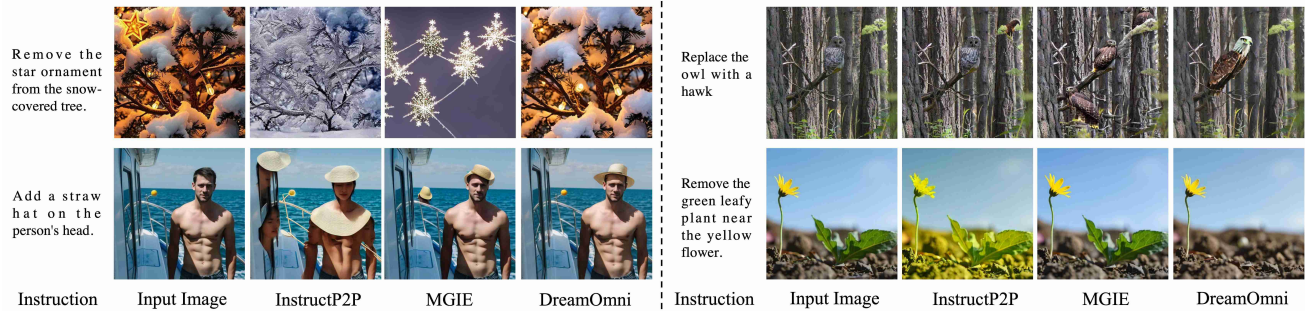**Evaluation on Instruction Editing.** We compare

Figure 8. Visual comparison on **instruction-based editing**. Our DreamOmni achieves more precise editing (including addition, removal, and replacement) compared to competitive methods such as MGIE [17] and InstructP2P [8].

Table 2. Quantitative comparison for inpainting & outpainting.

Model	Inpainting		Outpainting	
	FID↓	LPIPS↓	FID↓	LPIPS↓
SD-inpainting [44]	1.3522	0.1560	2.9179	0.2475
ControlNet-inpainting [62]	1.8393	0.1594	4.2337	0.2521
DreamOmni (Ours)	**0.8371**	**0.1203**	**1.6926**	**0.1995**

DreamOmni with competitive methods such as MGIE [17] and InstructP2P [8]. The visual results are presented in Fig. 8. We can see that DreamOmni performs more accurate edits, including addition, removal, and replacement. Specifically, our editing results exhibit superior consistency with minimal changes to non-edited areas and higher-quality generation of the edited content. This further validates that our synthetic data pipeline is an efficient and effective method for creating instruction-based editing datasets, enabling models to learn precise instruction-based editing. Moreover, the efficiency of our synthetic data pipeline allows for models to easily scale up diverse training data.

Evaluation on Drag Editing. We evaluate DreamOmni on our synthetic evaluation dataset. The visual results are shown in Fig. 9. (**1**) Compared to the target image, we can see that DreamOmni can accurately perform translation, rotation, and scaling drag edits. (**2**) for translation and scaling, DreamOmni can maintain the integrity of the dragged object. However, large-scale rotation operations are more challenging for DreamOmni, as they involve complex transformations of the object itself, which may lead to deformations of the edited object. (**3**) The results demonstrate the effectiveness of our synthetic data pipeline for drag editing, and encoding the drag point positions and displacement information as instruction inputs allows the model to learn precise drag edits (shown in Fig. 2).

5. Conclusion

Current T2I foundation models lack a unified framework and training for downstream tasks, such as image editing. To address this, we introduce DreamOmni, a unified model for T2I generation and editing. We evaluate the frameworks
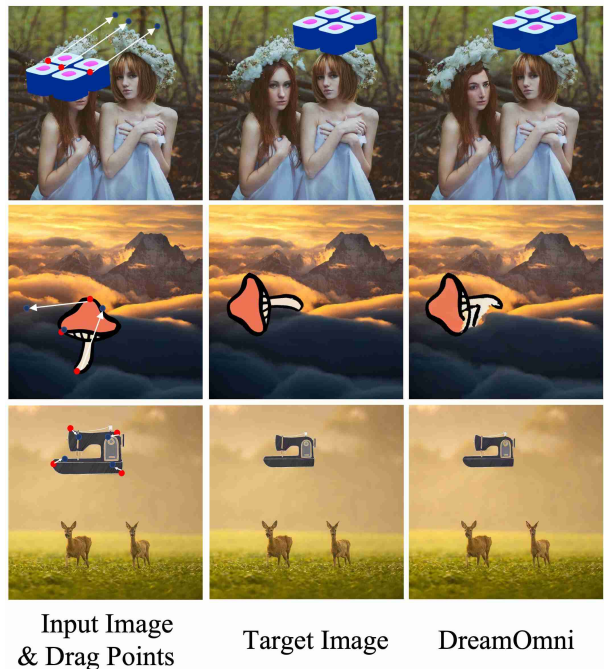

Figure 9. Visual results of **drag editing**: DreamOmni accurately performs translation, rotation, and scaling edits.

of existing models under fair settings and consider the specific needs of various editing tasks. From this analysis, we develop a framework that integrates T2I with various editing tasks. Besides, a challenge in training editing models is the creation of high-quality, large-scale editing data, which is inefficient. To overcome this, we designed a synthetic collage data pipeline capable of efficiently generating quite a few precise, high-quality editing data. Moreover, the pipeline enhances the model's generation accuracy in text, position, quantity, color, and geometry. By jointly training on T2I and multi-task synthetic data, we develop a native, unified model for both image generation and editing. T2I training strengthens the model's grasp of specific concepts and improves generation quality, while editing training enables it to handle the requirements of editing tasks.

References

[1] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *NeurIPS*, 2021. 2

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 3

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 3

[4] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, 2023. 3

[5] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021. 2

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

[7] Manuel Brack, Felix Friedrich, Katharina Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. *arXiv preprint arXiv:2311.16711*, 2023. 3

[8] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2, 3, 4, 8

[9] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023. 3

[10] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *ICLR*, 2024. 3, 5, 6

[11] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 5

[12] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 3

[13] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xi-

aofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 3

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021. 2

[15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 1, 3, 4, 5, 6, 7

[16] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *CVPR*, 2023. 3

[17] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *ICLR*, 2024. 8

[18] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, et al. Instructdiffusion: A generalist modeling interface for vision tasks. *arXiv preprint arXiv:2309.03895*, 2023. 3

[19] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS*, 36, 2024. 5, 7

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014. 2

[21] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022. 2

[22] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2

[24] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 2022. 2

[25] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. *CVPR*, 2024. 3

[26] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[27] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 1

[28] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pretrained subject representation for controllable text-to-image generation and editing. *NeurIPS*, 2024. 3, 4, 7

[29] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet ++: Improving conditional controls with efficient consistency feedback. In *ECCV*, 2025. 3

[30] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, 2024. 3

[31] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 3

[32] Pengyang Ling, Lin Chen, Pan Zhang, Huaian Chen, Yi Jin, and Jinjin Zheng. Freedrag: Feature dragging for reliable point-based image editing. In *CVPR*, 2024. 3

[33] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 5

[34] Hayk Manukyan, Andranik Sargsyan, Barsegh Atanyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Hd-painter: high-resolution and prompt-faithful text-guided image inpainting with diffusion models. *arXiv preprint arXiv:2312.14091*, 2023. 3

[35] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021. 2

[36] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023. 3

[37] 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. 3

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2, 3, 4

[39] 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. *NeurIPS*, 2023. 1, 3, 4, 5, 6

[40] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 3

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[42] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 2020. 4

[43] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 3

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 4, 5, 6, 7, 8

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[46] 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 *CVPR*, 2023. 3

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 1, 3

[48] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022. 5

[49] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. *arXiv preprint arXiv:2311.10089*, 2023. 3

[50] Yujun Shi, Jun Hao Liew, Hanshu Yan, Vincent YF Tan, and Jiashi Feng. Instadrag: Lightning fast and accurate drag-based image editing emerging from videos. *arXiv preprint arXiv:2405.13722*, 2024. 3, 5

[51] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *CVPR*, 2024. 3

[52] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[53] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[54] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023. 3

[55] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *CVPR*, 2023. 3

[56] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 5

[57] Bin Xia, Shiyin Wang, Yingfan Tao, Yitong Wang, and Jiaya Jia. Llmga: Multimodal large language model based generation assistant. In *European Conference on Computer Vision*, 2025. 3

[58] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. 3

[59] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *CVPR*, 2023. 3

[60] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3, 4, 7

[61] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *NeurIPS*, 2024. 3

[62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 5, 6, 7, 8

[63] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *NeurIPS*, 2023. 3