OmniGen: Unified Image Generation

Shitao Xiao*, **Yueze Wang***, **Junjie Zhou***, **Huaying Yuan***,
Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, Zheng Liu†
Beijing Academy of Artificial Intelligence
{stxiao, yzwang}@baai.ac.cn, zhengliu1026@gmail.com

Figure 1: OmniGen demonstrates the capability to perform various image generation tasks within a single framework. Additionally, it possesses reasoning abilities and in-context learning capabilities.

*Co-first authors
†Corresponding authors

Preprint. Under review.

Abstract

The emergence of Large Language Models (LLMs) has unified language generation tasks and revolutionized human-machine interaction. However, in the realm of image generation, a unified model capable of handling various tasks within a single framework remains largely unexplored. In this work, we introduce OmniGen, a new diffusion model for unified image generation. Unlike popular diffusion models (e.g., Stable Diffusion), OmniGen no longer requires additional modules such as ControlNet or IP-Adapter to process diverse control conditions. OmniGen is characterized by the following features: 1) **Unification**: OmniGen not only demonstrates text-to-image generation capabilities but also inherently supports various downstream tasks, such as image editing, subject-driven generation, and visual-conditional generation. Additionally, OmniGen can handle classic computer vision tasks by transforming them into image generation tasks, such as edge detection and human pose recognition. 2) **Simplicity**: The architecture of OmniGen is highly simplified, eliminating the need for additional text encoders. Moreover, it is more user-friendly compared to existing diffusion models, enabling complex tasks to be accomplished through instructions without the need and cost for extra preprocessing steps (e.g., human pose estimation), thereby significantly simplifying the workflow of image generation. 3) **Knowledge Transfer**: Benefit from learning in a unified format, OmniGen effectively transfers knowledge across different tasks, manages unseen tasks and domains, and exhibits novel capabilities. We also explore the model's reasoning capabilities and potential applications of chain-of-thought mechanism. This work represents the first attempt at a general-purpose image generation model, and there remain several unresolved issues. We will open-source the related resources at https://github.com/VectorSpaceLab/OmniGen to foster advancements in this field.

1 Introduction

The pursuit of Artificial General Intelligence (AGI) has intensified the demand for generative foundation models capable of handling a wide variety of tasks within a single framework. In the field of Natural Language Processing (NLP), Large Language Models (LLMs) have become exemplary in achieving this goal, demonstrating remarkable versatility across numerous language tasks such as question answering, text summarization, and code generation.

However, the field of visual generation has yet to reveal a counterpart that mirrors the universality of LLMs. Current image generation models have demonstrated proficiency in specialized tasks. For instance, in the text-to-image generation filed, state-of-the-art models such as the Stable Diffusion series [56; 52; 13], DALL-E [55], and Imagen [26] have made significant strides. Meanwhile, many efforts have been proposed to extend and optimize the capabilities of diffusion models for specific tasks. Models like ControlNet [73] and T2i-Adapter [45] design an additional network plugged into the text-to-image diffusion model to support visual conditions. InstructPix2Pix [4] is trained on a comprehensive dataset tailored for image editing tasks. Despite their strengths, those models are limited by their task-specific nature and do not exhibit the comprehensive perceptual understanding and generative capabilities required for a universal model in visual generation.

Is it possible to address various image generation tasks, such as text-to-image, image editing, controllable generation, and image restoration, within a single diffusion framework, akin to how GPT handles language tasks? If a universal model is available, the need for training additional modules (e.g., ControlNet, IP-Adapter, T2I-Adapter) in practical applications can be eliminated. Motivated by this potential, we explore a unified framework for image generation, named OmniGen.

Unlike popular diffusion models, OmniGen features a very concise structure, comprising only two main components: a VAE and a transformer model, without any additional encoders. OmniGen supports arbitrarily interleaved text and image inputs as conditions to guide image generation, rather than text-only or image-only conditions. To train a robust unified model, we construct the first large-scale unified image generation dataset X2I, which unifies various tasks into one format. Additionally, we incorporate several classic computer vision tasks such as human pose estimation, edge detection, and image deblurring, thereby extending the model's capability boundaries and

enhancing its proficiency in complex image generation tasks. We evaluate our model on multiple benchmarks, demonstrating its competitive text-to-image generation capabilities compared to existing models. Furthermore, our model inherently supports various image generation tasks, such as image editing, visual conditional generation, and subject-driven generation, which are beyond the reach of current diffusion models. Remarkably, the design of OmniGen allows for robust transfer learning across different scenarios, facilitating the handling of previously unseen tasks and domains, as well as giving birth to emerging abilities. Our contributions are summarized below:

- We introduce OmniGen, a unified model for image generation that excels in multiple domains. OmniGen demonstrates competitive text-to-image generation capabilities and inherently supports a variety of downstream tasks such as controllable image generation and subject-driven generation. Furthermore, it is capable of performing classic computer vision tasks. To the best of our knowledge, OmniGen is the first image generation model to achieve such a comprehensive level of functionality.
- We construct a comprehensive image generation dataset named X2I, which stands for "anything to image". This dataset includes a wide range of image generation tasks, all standardized into a unified format.
- By unified training on the multi-task dataset, OmniGen can apply learned knowledge to tackle unseen tasks and domains, as well as exhibit new capabilities. Additionally, OmniGen shows a degree of reasoning capability.

The remainder of this paper is organized as follows: Section 2 details the model architecture, while Section 3 describes the dataset construction. Section 4 presents the model's performance on various image generation tasks. In Section 5, we analyze the model's emerging capabilities and reasoning abilities, and also explore the potential applications of CoT in image generation. Section 6 discusses the current limitations of the model. Section 7 reviews related work.

2 OmniGen

In this section, we present the details of OmniGen framework, including the model architecture and training method.

2.1 Model Design

Principles. Current diffusion models are typically limited to common text-to-image tasks and can not perform a broader range of downstream image-generation tasks. To achieve real-world applications, users often need to design and integrate additional network structures to extend the capabilities of diffusion models, making the models highly cumbersome. Even worse, these additional parameter networks are usually task-specific and can not be reused for other tasks, unless more networks are designed and trained for different functions. To circumvent these issues, the design principles of OmniGen are as follows: 1). Universality: accepting any form of image and text inputs for various tasks; 2). Conciseness, avoiding overly complex structural designs and numerous additional components.

Network Architecture. As illustrated in Figure 2, the OmniGen framework adopts an architecture comprised of a Variational Autoencoder (VAE) [28] and a pre-trained large transformer model. Specifically, VAE extracts continuous visual features from images, while the transformer model generates images based on input conditions. In this paper, we use the VAE from SDXL [52] and freeze it during training. We use Phi-3 [1] to initialize the transformer model, inheriting its excellent text processing capabilities. Unlike state-of-the-art diffusion models that require additional encoders to pre-process conditional information (such as clip text encoder and image encoder), OmniGen inherently encodes conditional information by itself, significantly simplifying the pipeline. Furthermore, OmniGen jointly models text and images within a single model, rather than independently modeling different input conditions with separate encoders as in existing works [67; 68; 70; 63; 9] which lacks interaction between different modality conditions.

Input Format. The input to the model can be multimodal interleaved text and images in free form. We utilize the tokenizer of Phi-3 to process text without any modifications. For images, we firstly employ a VAE with a simple linear layer to extract latent representations. Then, they are flattened

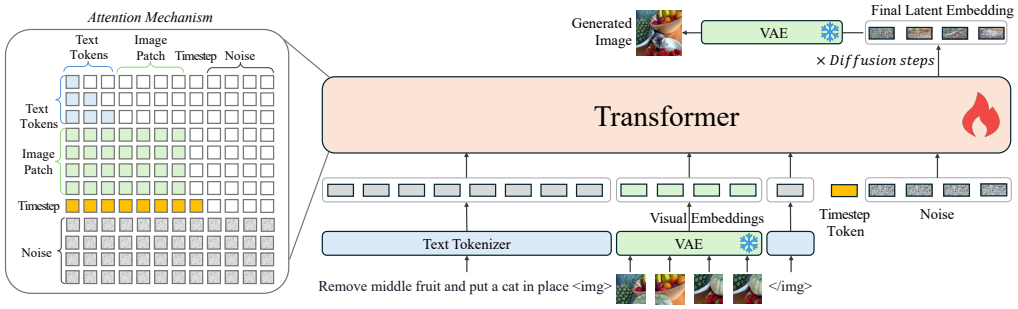


Figure 2: The framework of OmniGen. Texts are tokenized into tokens, while input images are transformed into embedding via VAE. OmniGen can accept free-form multi-modal prompts and generate images through the rectified flow approach.

into a sequence of visual tokens by linearly embedding each patch in the latent space. Following [50], we apply standard frequency-based positional embeddings to input visual tokens, and use the same method as SD3 [13] to process images with varying aspect ratios. In addition, we encapsulate each image sequence with two special tokens: "<img>" and "</img>" before inserting it into the text tokens sequence. We also add the timestep embedding [50] at the end of the input sequence.

Attention Mechanism. Different from text, which can be decomposed into discrete tokens to model, we argue that images should be modeled as a whole. Therefore, we modify the common causal attention mechanism in LLM, integrating it with the bidirectional attention as illustrated in Figure 2. Specifically, we apply causal attention to each element in the sequence, but apply bidirectional attention within each image sequence. This allows each patch to pay attention to other patches within the same image, while ensuring that each image can only attend to other images or text sequences that have appeared previously. .

Inference. During inference, we randomly sample a Gaussian noise and then apply the flow matching method to predict the target velocity, iterating multiple steps to obtain the final latent representation. Finally, we use a VAE to decode the latent representation into the predicted image. The default inference step is set to 50. Thanks to the attention mechanism, OmniGen can accelerate inference like LLMs by using kv-cache: storing previous and current key and value states of the input conditions on the GPU to compute attention without redundant computations.

2.2 Training Strategy

Train objective. In this work, we use rectified flow [41] to optimize the parameters of model. Different from DDPM [25], flow matching conducts the forward process by linearly interpolating between noise and data in a straight line. At the step t, $\mathbf{x}_t$ is defined as

$$\mathbf{x}_t = t\mathbf{x} + (1-t)\boldsymbol{\epsilon},$$

where $\mathbf{x}$ is the original data, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0,1)$ is the Gaussian noise. The model is trained to directly regress the target velocity given the noised data $\mathbf{x}_t$, timestep t, and condition information c. Specifically, the objective is to minimize the mean squared error loss:

$$\mathcal{L} = \mathbb{E}\left[||(\mathbf{x}-\boldsymbol{\epsilon}) - v_\theta(\mathbf{x}_t,\ t,\ c)||^2\right]. \tag{1}$$

For image editing tasks, the objective is to modify specific regions of the input image while keeping other areas unchanged. Therefore, the difference between the input image and the target image is often small, which allows the model to learn an unexpected shortcut: simply copying the input image as the output to make the related training loss very low. To mitigate this phenomenon, we amplify the loss in the regions of the image where changes occur. More specifically, we calculate the loss weights for each region based on these latent representations of input image $\mathbf{x}'$ and target image $\mathbf{x}$:

$$w_{i,j} = \begin{cases} 1 & \text{if } \mathbf{x}_{i,j} = \mathbf{x'}_{i,j} \\ \frac{1}{||\mathbf{x}-\mathbf{x}'||^2} & \text{if } \mathbf{x}_{i,j} \neq \mathbf{x'}_{i,j} \end{cases} \tag{2}$$

Consequently, regions with alterations are assigned higher weights than those without changes, guiding the model to focus on the areas to be modified.

Training Pipeline. Following previous work [13; 18; 6], we gradually increase the image resolution during the training process. Low resolution is data-efficient, while high resolution can enhance the aesthetic quality of the generated images. Detailed information for each training stage is presented in Table 2.2. We adopt the AdamW [42] with $\beta = (0.9, 0.999)$ as the optimizer. All experiments are conducted on 104 A800 GPUs.

Stage	Image Resolution	Training Steps (K)	Batch Size	Learning Rate
1	256×256	500	1040	1e-4
2	512×512	300	520	1e-4
3	1024×1024	100	208	4e-5
4	2240×2240	30	104	2e-5
5	Multiple	80	104	2e-5

Table 1: Detailed information about every OmniGen training stage.

3 X2I Dataset

To achieve robust multi-task processing capabilities, it is essential to train models on large-scale and diverse datasets. However, in the field of image generation, a readily available large-scale and diverse dataset has yet to emerge. In this work, we have constructed a large-scale unified image generation dataset for the first time, which we refer to as the **X2I** dataset, meaning "anything to image". We have converted these data into a unified format, and Figure 3 presents some examples from the X2I dataset. The entire dataset comprises approximately 0.1 billion images. We will provide a detailed description of the composition of this dataset in the following sections.

3.1 Text to Image

The input for this subset of data is plain text. We have obtained multiple open-source datasets from various sources: Recap-DataComp [35](a subset of 56M images), SAM-LLaVA [6], ShareGPT4V [7], LAION-Aesthetic [58](a subset of 4M images), ALLaVA-4V [5], DOCCI [47], DenseFusion [36] and JourneyDB [60]. While these datasets are large in quantity, their image quality is not always high enough. In the early stages of training, we use them to learn a broad range of image-text matching relationships and diverse knowledge. After stage 3, we utilize our internal collection of 16 million high-quality images to enhance the aesthetic quality of the generated images. A lot of studies [13; 6] have demonstrated that synthetic detailed captions can greatly improve text-to-image models trained at scale. Therefore, we use the InternVL2 [11] to create synthetic annotations for internal data and LAION-Aesthetic (the other datasets come with detailed text descriptions and do not require further annotation).

3.2 Multi-modal to Image

Different from most existing diffusion models, our model can accept more general and flexible multimodal instruction as conditions to guide the generation of images.

3.2.1 Common Mixed-modal Prompts

The input of this portion of data is arbitrarily interleaved text and images. We collect the data from multiple tasks and sources: image editing (SEED-Data-Edit [19], MagicBrush [72], and InstructPix2Pix [4]), human motion (Something-Something [23]), virtual try-on (HR-VITON [31] and FashionTryon [75]), and style transfer (stylebooth [24]). We standardize all tasks into the input-output pair format as shown in Figure 3-(b).

The issue of utilizing additional visual conditions for finer-grained spatial control has garnered widespread attention[73; 33]. We employ the MultiGen [53] dataset to learn this function, and select six representative visual conditions: Canny, HED, Depth, Skeleton, Bounding Box, and segmentation.

Figure 3: Examples of our training data for the OmniGen model. We standardized the input of all tasks into an arbitrarily interleaved image-text sequence format, used as the model's prompt. The placeholder |image_i| represents the position of the i-th image within the prompt.

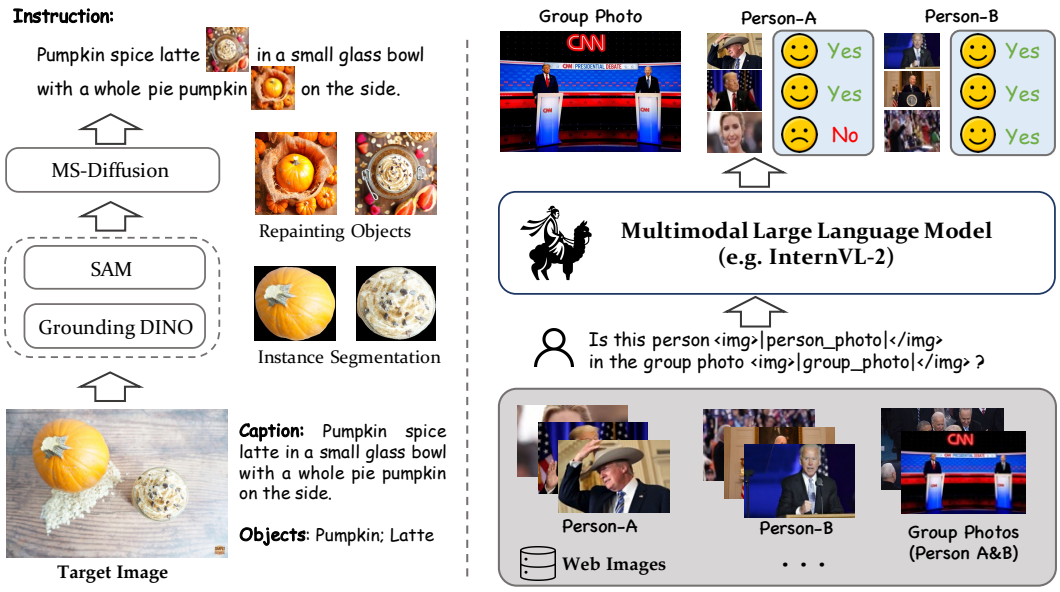


Figure 4: **(a) Illustration of the construction process for the GRIT-Entity dataset.** We used instance segmentation and repainting methods to acquire a large volume of data. **(b) Illustration of the cross-verification strategy used in constructing our web images dataset.** For a group photo of Person-A and Person-B, we sampled several images from individual photos of Person-A and Person-B and asked MLLM whether they appear in the group photo. A group photo is retained only if the "Yes" ratio for both Person-A and Person-B meets a specific threshold. The individual images marked as "Yes" are then used to construct data pairs with the corresponding group image.

These types of tasks take text prompts and specific visual conditions (such as segmentation maps, and human pose maps) as multi-modal inputs, then generate new images that comply with the text and image conditions.

3.2.2 Subject-driven Image Generation

We constructed both a large-scale foundational dataset (GRIT-Entity dataset) and a high-quality advanced dataset (Web Images dataset) for subject-driven image generation. For the GRIT-Entity dataset, we leveraged the GRIT dataset [51], which annotates object names within images. Using these annotations, we applied the Grounding DINO model [40] for text-to-bounding-box grounding. Based on the bounding boxes, we employed SAM [29] to segment the cropped images, obtaining object masks. We further used the MS-Diffusion model [64] to repaint the object images, enhancing data quality. The process of data construction and the final instruction format are illustrated in Figure 4-(a). Through this method, we acquired 6 million pairs.

Although the GRIT-based approach provides a substantial amount of data, the input data extracted directly from original images can lead the model to fall into simple copy-paste patterns. To fully unleash the subject-driven image generation capability of OmniGen, we constructed a high-quality web images training dataset using natural images of well-known individuals. First, we sampled 20 million Alt-text entries from the Datacomp dataset [14] and used spaCy[3] for named entity recognition. We selected the most frequently occurring names and employed GPT-4o to filter out real, notable individuals, resulting in 2,000 names. Furthermore, we expanded the initial 2,000 names by including closely related individuals, resulting in approximately 10,000 name pairs. We then scraped images of these individuals and pairs from search engines. Due to the noise in web images, where scraped images may not contain the specified individuals, we designed a cross-verification strategy using InternVL [11] to filter single and group images, as detailed in Figure 4-(b). The retained single and group images were then captioned with details such as attire and actions. Through additional

[3]https://github.com/explosion/spaCy

instruction-based annotations, we successfully constructed a dataset of 533,000 image pairs. We present some examples in Figure 3-(c).

3.2.3 Computer Vision Tasks

We introduce classic computer vision tasks to enhance the image generation capabilities of the model. For low-level vision tasks (low-light image enhancement [66], deraining [71], deblurring [46], inpainting [53], outpainting [53] and colorization [58]), where the annotation itself is an image, we only add text instructions, which were randomly sampled from instructions generated by GPT-4o. For high-level tasks, we choose to represent all annotations as images. We used LAION [58] as the source image and annotations from [53] as the target to construct image pairs (such as source image and its human pose mapping). The annotations include human pose, depth mapping, canny, and segmentation. Additionally, we also use several datasets for referring image segmentation, including RefCOCO [27], ADE20k [76], and ReasonSeg [30]. As shown in Figure 3-(c), the input is the source image and a natural language expression, the output is an image with the corresponding object highlighted in blue.

The purpose of constructing these datasets is not merely to endow the model with these functionalities. We also aim to transfer the knowledge acquired from these traditional computer vision tasks to image generation tasks, thereby achieving more sophisticated image generation capabilities. Our experiments have also demonstrated that multi-task learning enables the model to exhibit emergent abilities.

3.3 Few-shot to Image

We constructed a few-shot to image dataset to stimulate the model's in-context learning capabilities. Specifically, for each task described in the preceding sections, we randomly selected a few examples and combined the original input with these examples to form new inputs. The specific data format can be referenced in Figure 3-(e). Due to limitations in training resources, we opted to use only one example to enhance training efficiency.

4 Experimental Results

In this section, we show the results of OmniGen in image generation tasks and traditional vision tasks.

4.1 Image Generation

4.1.1 Qualitative Results

Figure 5 shows the results of the text-to-image task. It can be observed that OmniGen effectively follows the textual descriptions to generate images with arbitrary aspect ratios.

Figure 6 presents the outcomes of the subject-driven generation task. Our model can extract the required objects from the given reference images and generate new images accordingly. Furthermore, when the reference image contains multiple objects, the model can directly select the needed objects based on textual instructions (e.g., the cat in the figure) without requiring additional preprocessing steps such as image cropping or face recognition.

Figure 7 summarizes the results of other image generation tasks, demonstrating that the model can handle various downstream tasks based on multi-modal instructions.

4.1.2 Text to Image

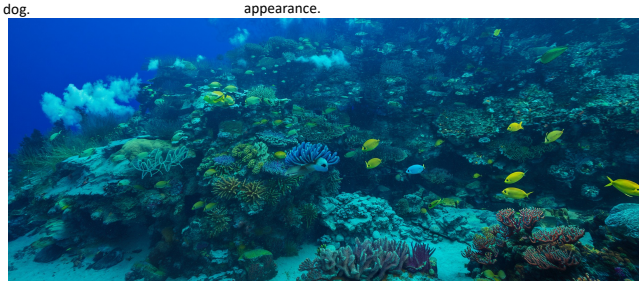
Following [13], we evaluate text-to-image generation capability of OmniGen on the GenEval [22] benchmark. We compared the performance of our model with the reported results of other popular image generation models, as summarized in Table 2. Surprisingly, our model achieved similar performance compared to the current state-of-the-art diffusion models, such as SD3, which underscores the effectiveness of our framework. The GenEval benchmark does not reflect the aesthetic quality of images, we will leave this aspect for future evaluation.

A young woman stands on the platform of a railway station in spring. She wore a long blue–gray windbreaker and white shirt. Her darkbrown hair was tied in a low ponytail, a few strands flying in the wind. Her eyes were full of expectation, and the sun shone on her warm face.

An extreme close-up of an gray-haired man with a beard in his 60s, he is deep in thought pondering the history of the universe as he sits at a cafe in Paris, his eyes focus on people offscreen as they walk as he sits mostly motionless, he is dressed in a wool coat suit coat with a button-down shirt, he wears a brown beret and glasses and has a very professorial appearance.

A portrait of a statue of the Egyptian god Anubis wearing aviator goggles, white t-shirt and leather jacket. The skyline of Los Angeles at night can be seen in the background.

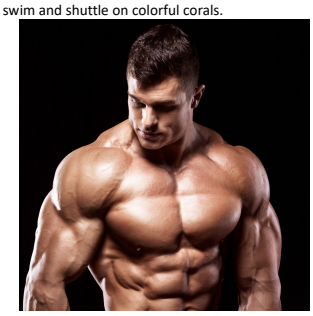
On the left, a golden retriever sits with its tongue out. On the right, a tabby cat lounges lazily, stretching its legs out and looking towards the dog.



In the deep sea, all kinds of marine life swim and shuttle on colorful corals.

A bull's head warrior in boots and armor, holding an epee, excited and majestic.

Leaves made of ice



A muscular man with prominent muscular lines.

A cheesburger surfing the vibe wave at night

A bright red rose, petals scattered with some water droplets, crystal clear, close-up.

Figure 5: Examples for text-to-image task. OmniGen can generate images with arbitrary aspect ratios.

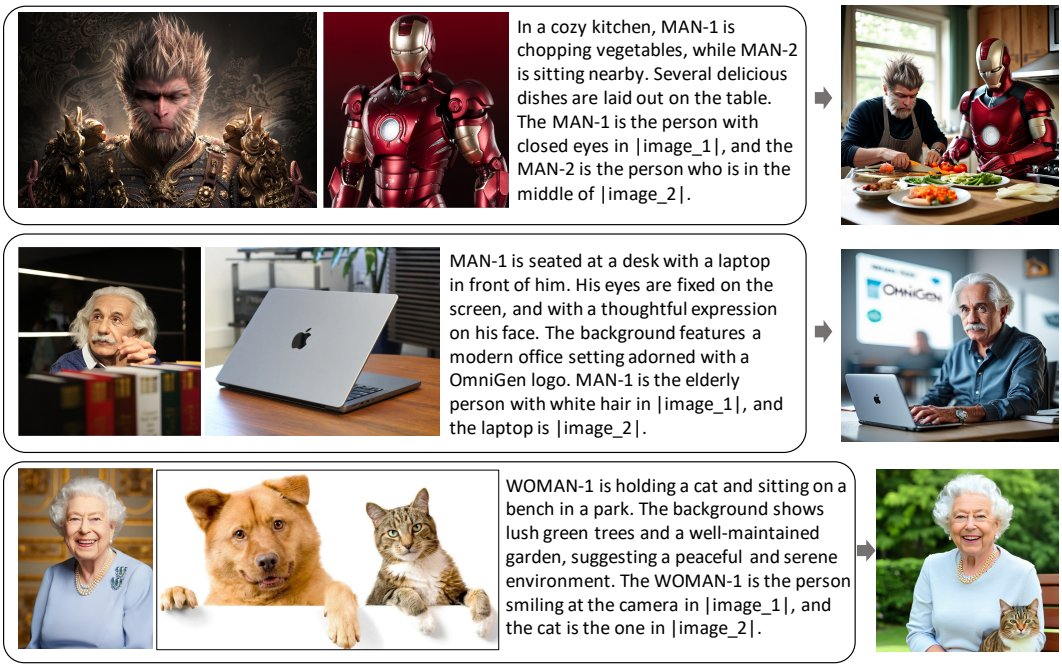


Figure 6: The results of subject-driven generation. OmniGen can generate new image based on the objects from reference images. When the reference image contains multiple objects, OmniGen can automatically identify the required objects based on textual instructions.

Figure 7: The results of OmniGen in different image generation tasks.

Model	Params	Data	Overall	Single object	Two object	Counting	Colors	Position	Attribute binding
SDv1.5	0.9B+0.1B*	–	0.43	0.97	0.38	0.35	0.76	0.04	0.06
SDv2.1	0.9B+0.4B*	–	0.50	0.98	0.51	0.44	**0.85**	0.07	0.17
SD-XL	2.6B+0.8B*	–	0.55	0.98	0.74	0.39	**0.85**	0.15	0.23
DALLE-2	3.5B+1.0B*	–	0.52	0.94	0.66	0.49	0.77	0.10	0.19
DALLE-3	–	–	0.67	0.96	**0.87**	0.47	0.83	**0.43**	0.45
IF-XL	5.5B+4.7B*	1.2B	0.61	0.97	0.74	**0.66**	0.81	0.13	0.35
SD3	8.0B+4.7B*	1.0B	0.68	0.98	0.84	**0.66**	0.74	0.40	0.43
OmniGen	3.8B	0.1B	**0.70**	**0.99**	0.86	0.62	**0.85**	0.32	**0.54**

Table 2: Results of GenEval Benchmark. * means the parameter of the frozen text encoder. Our model achieves similar performance with a relatively small scale and limited data.

Notably, our model has only 3.8 billion parameters, whereas the SD3 model has a total of 12.7 billion parameters (more than three times that of ours). Current diffusion models typically adopt an encoder-decoder architecture, utilizing an additional encoder to encode textual conditions (this text encoder alone is larger than our entire model). In contrast, our model architecture is significantly simplified, eliminating the cost of an additional text encoder, thereby greatly enhancing the efficiency of parameter utilization. Besides, we employed only 0.1 billion image data, whereas SD3 used over 1 billion (more than ten times that of ours), highlighting the role of multitask data **X2I** in enhancing text-to-image capabilities.

4.1.3 Image Edit

Model	CLIP-I↑	CLIP-T↑	DINO↑
InstructPix2Pix [4]	0.834	0.219	0.762
MagicBrush [72]	0.838	0.222	0.776
PnP [62]	0.521	0.089	0.153
Null-Text Inv. [44]	0.761	**0.236**	0.678
EMU-Edit [59]	**0.859**	0.231	**0.819**
OmniGen	0.836	0.233	0.804

Table 3: Results on EMU-Edit test data. As a universal model, OmniGen demonstrates performance comparable to that of the best proprietary models.

We compare OmniGen with other state-of-the-art image editing models on EMU-Edit [59] dataset, which includes seven different operations: background alteration, comprehensive image changes, style alteration, object removal, object addition, localized modifications, and color/texture alterations. We measure three metrics: 1) CLIP-I: CLIP image similarity between the source image and output image; 2) DINO: DINO similarity between the source image and output image; and 3) CLIP-T: CLIP text-image similarity between edited image and target caption. DINO and CLIP-I similarity scores measure the model's ability to preserve elements from the source image, while CLIP-T measures how well the model followed the instructions. As shown in Table 2, our model significantly outperforms InstructPix2Pix [4], and exhibits comparable performance to the current state-of-the-art model: EMU-Edit [59].

4.1.4 DreamBooth

We evaluate the single-entity subject-driven generation capability on DreamBench [57]. The DreamBench contains 750 prompts for 30 subjects (e.g., dog and toy). For each prompt, we generate 4 images, resulting in a comprehensive evaluation set of 3,000 images. Following Kosmos-G [49], we only select one image as input from the 4-7 provided images for each subject. We adopted DINO and CLIP-I from DreamBooth to assess subject fidelity, and CLIP-T for text fidelity. All results are summarized in Table 4. Compared to methods based on fine-tuning, our approach maintains a comparable level of text fidelity while better preserving the subject from the source image. Compared with models without fine-tuning, OmniGen significantly outperforms both Re-Imagen and Kosmos-G, and demonstrates superior subject fidelity relative to SuTI.

Model	DINO↑	CLIP-I↑	CLIP-T↑
Fine-Tuning			
Textual Inversion [15]	0.569	0.780	0.255
DreamBooth [57]	0.668	0.803	**0.305**
BLIP-Diffusion [32]	0.670	0.805	0.302
Test Time Tuning Free			
Re-Imagen [8]	0.600	0.740	0.270
SuTI [10]	<u>0.741</u>	<u>0.819</u>	<u>0.304</u>
Kosmos-G [49]	0.694	**0.847**	0.287
OmniGen	**0.801**	**0.847**	0.301

Table 4: Results on DreamBench. OmniGen achieves better subject fidelity and text fidelity.

4.1.5 Visual Conditional Controls

Image-based prompts can provide detailed spatial conditioning controls for diffusion models. To evaluate this ability of OmniGen, we use the dataset and script from [33]. This benchmark includes ADE20K test dataset for segmentation mask condition, and evaluation split of MultiGen-20M for canny edge map, hed edge map, and depth map condition. For each condition, the controllability is evaluated by measuring the similarity between the input conditions and the extracted conditions from generated images of diffusion models. The experimental results are shown in Table 5. We can find that our model achieves optimal results on segmentation mask and hed edge map conditions, and obtains competitive results for canny edge map and depth map conditions.

	Seg. Mask (mIoU↑)	**Canny Edge** (F1 Score↑)	**Hed Edge** (SSIM↑)	**Depth Map** (RMSE↓)
T2I-Adapter [45]	12.61	23.65	-	48.40
Gligen [37]	23.78	26.94	0.5634	38.83
Uni-ControlNet [74]	19.39	27.32	0.6910	40.65
UniControl [53]	25.44	30.82	0.7969	39.18
ControlNet [73]	32.55	34.65	0.7621	35.90
ControlNet++ [33]	<u>43.64</u>	**37.04**	<u>0.8097</u>	**28.32**
OmniGen	**44.23**	<u>35.54</u>	**0.8237**	<u>28.54</u>

Table 5: Comparison with SOTA methods on controllable image generation. ↑ indicates higher result is better, while ↓ means lower is better.

4.2 Computer Vision Tasks

We present several qualitative results of computer vision tasks in Figure 8. OmniGen can handle various low-level vision tasks such as deraining, deblurring, and inpainting. In the bottom of Figure 8, we can see that OmniGen is also able to handle high-level tasks, such as human pose recognition and depth estimation.

So far, we have demonstrated that our model can generate images well based on visual conditions while also extracting visual conditions from raw images. This motivates us to ponder: can we directly use the model to generate new images based on a reference image in only one step, instead of first using a processor to extract spatial condition information and then inputting it into the model for generation? Surprisingly, even without having encountered such a task before, OmniGen handles it admirably. As shown in Figure 9, the existing workflow for ControlNet involves using a detector to extract spatial condition information from the reference image, and then loading the corresponding control module to model the spatial condition information for image generation, which requires multiple network components and operations. Now, only based on our model, we can directly input the reference image and text instruction (e.g., *Follow the depth mapping of this image <reference-image> to generate new image. The text description for new image is "..."*) to generate an image in only one step without any additional intermediate procedures. It can be observed that the model comprehends the instruction well; when tasked with using the human pose from the reference image,

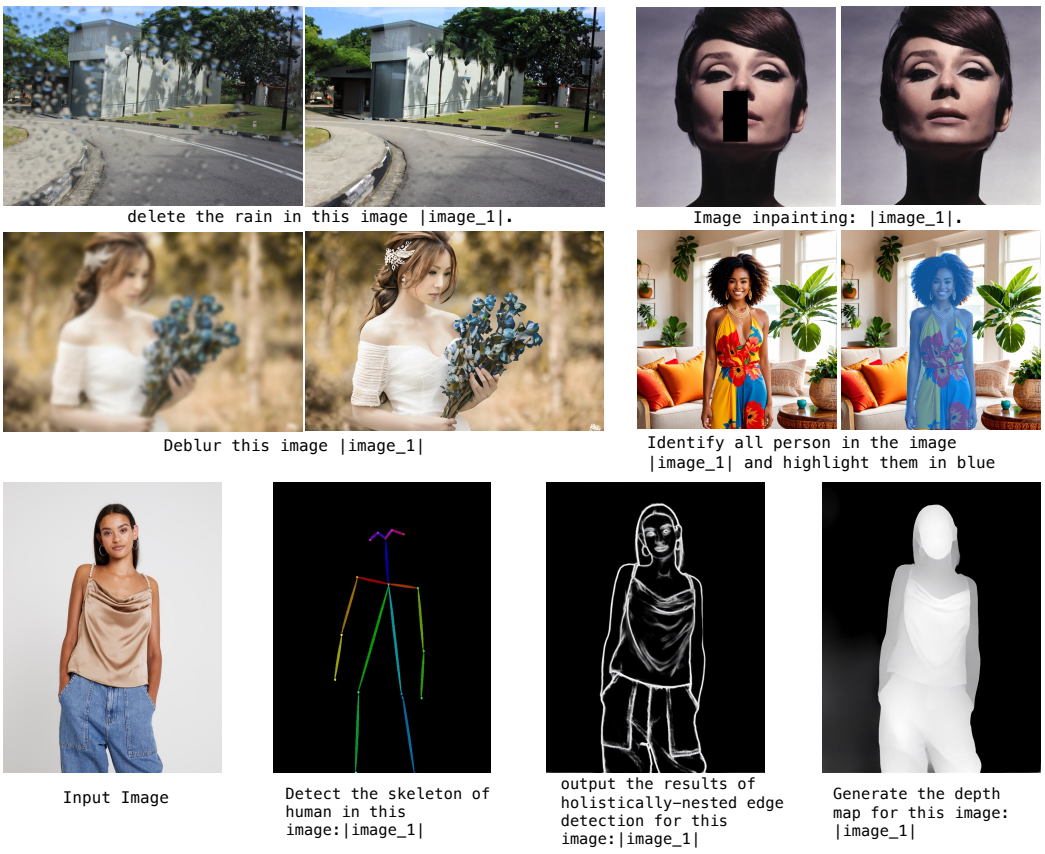


Figure 8: The results of OmniGen in traditional various vision tasks.

it perfectly replicates the human pose, and when using depth mapping, it retains more details, such as the folds in clothing.

5 Further Analysis

LLMs demonstrate remarkable generalization capabilities, achieving impressive performance in previously unseen tasks and domains. Furthermore, they can boost performance through mechanisms such as in-context learning and chain of thought. We observe similar functionalities in OmniGen as well, and present our findings in this section.

5.1 Emgerging Capabilities

By standardizing all tasks into a unified format and training on **X2I** dataset, OmniGen can acquire universal knowledge and allow knowledge transfer across different scenarios and tasks, thus enabling the generation capabilities on unseen tasks and domains. We illustrate several emerging capabilities using the following examples.

Task Composition. In real-world applications, user requirements often involve combinations of tasks. As shown in Figure 10-(a), our model is capable of simultaneously processing multiple instructions, including those for different tasks (*Image inpainting and change the color of hair to white*) as well as multiple instructions for the same task (*Add a sunglasses to the man's face, and change the color of clothes to blue*). These results highlight our model's versatility and potential for widespread adoption in the wild.

Implicit Combination of Tasks. In addition to explicit task combinations, our model is capable of performing multiple tasks implicitly through a single instruction. As demonstrated in Figure 9, upon receiving the input like "*follow the human pose/depth mapping to generate the image: ...*", our model

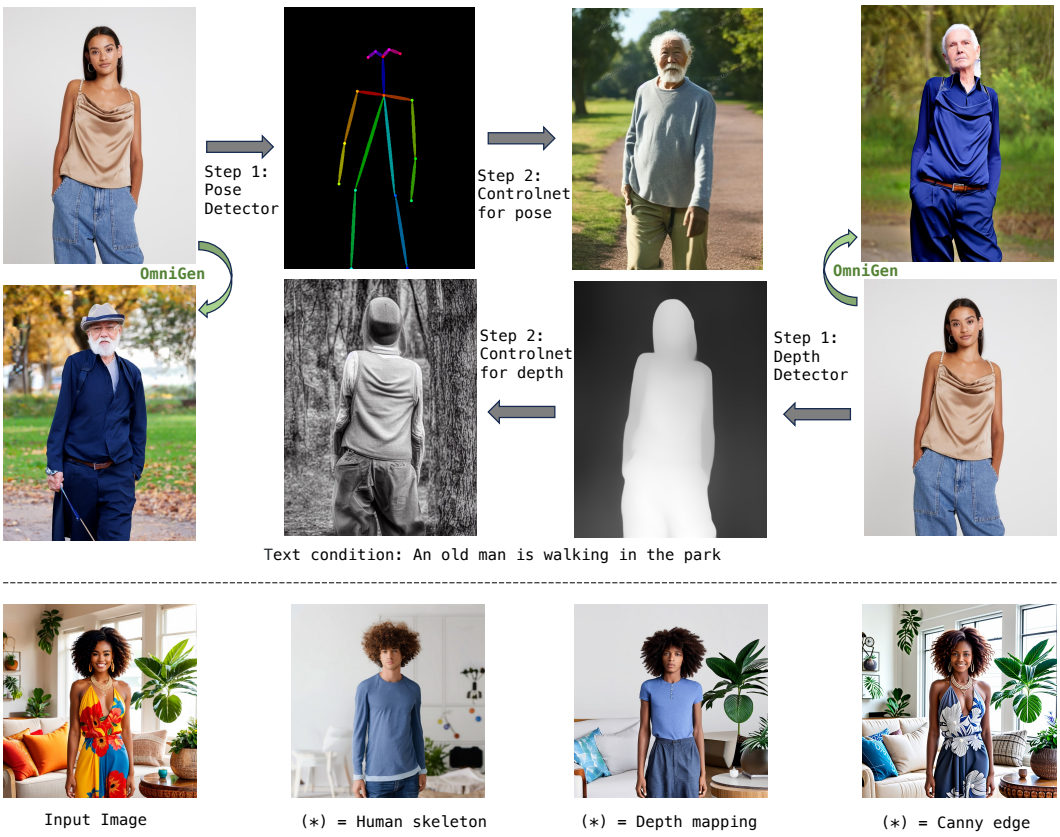


Figure 9: **Top**: Comparison with ControlNet. ControlNet involves two steps: first, using the corresponding detector to extract information from the reference image, and then loading the appropriate ControlNet module to process the visual conditions. In contrast, OmniGen completes the entire task in a single step, achieving superior image quality. **Bottom**: Results using different visual conditions. It can be observed that more detailed conditions lead to a higher similarity between the model's output and the reference image.

can extract the relevant conditional information (such as human pose, depth mapping, etc.) from the reference image and generate a new image based on the captured condition. This process is implicit, with all processing completed internally within the model, thus only requiring the user to input a simple command. This negates the need for explicit conditional extraction using other models prior to input into the diffusion process, as is necessary with existing systems like ControlNet [73].

In-context Learning for Unseen Tasks and Domains. As illustrated in Figure 10-(b), by providing an example, the model is capable of successfully completing a novel task: generating images based on provided scribble data, which is not encountered during training. To explore whether in-context learning can boost existing abilities on new domains, we show several examples from the FSS [34] dataset, which contains objects that have never been seen or annotated in previous datasets. In the left of Figure 10-(c), we can see that OmniGen is not familiar with the concepts of "pencil sharpeners" and "chess queens", and it cannot identify them from images. However, when provided with an example, the model is capable of making accurate predictions, demonstrating that in-context learning can enhance the model's generalization ability across different domains.

End-to-end Workflow. Users typically need to load multiple models and perform multi-step processing to ultimately generate a satisfactory image, making the workflow very cumbersome and costly. This complexity has led to the development of open-source tools and pipelines like ComfyUI[4]. Our model possesses both excellent multi-modal understanding and image generation capabilities, enabling it to complete a lot of tasks without relying on external models, thereby

[4]https://github.com/comfyanonymous/ComfyUI

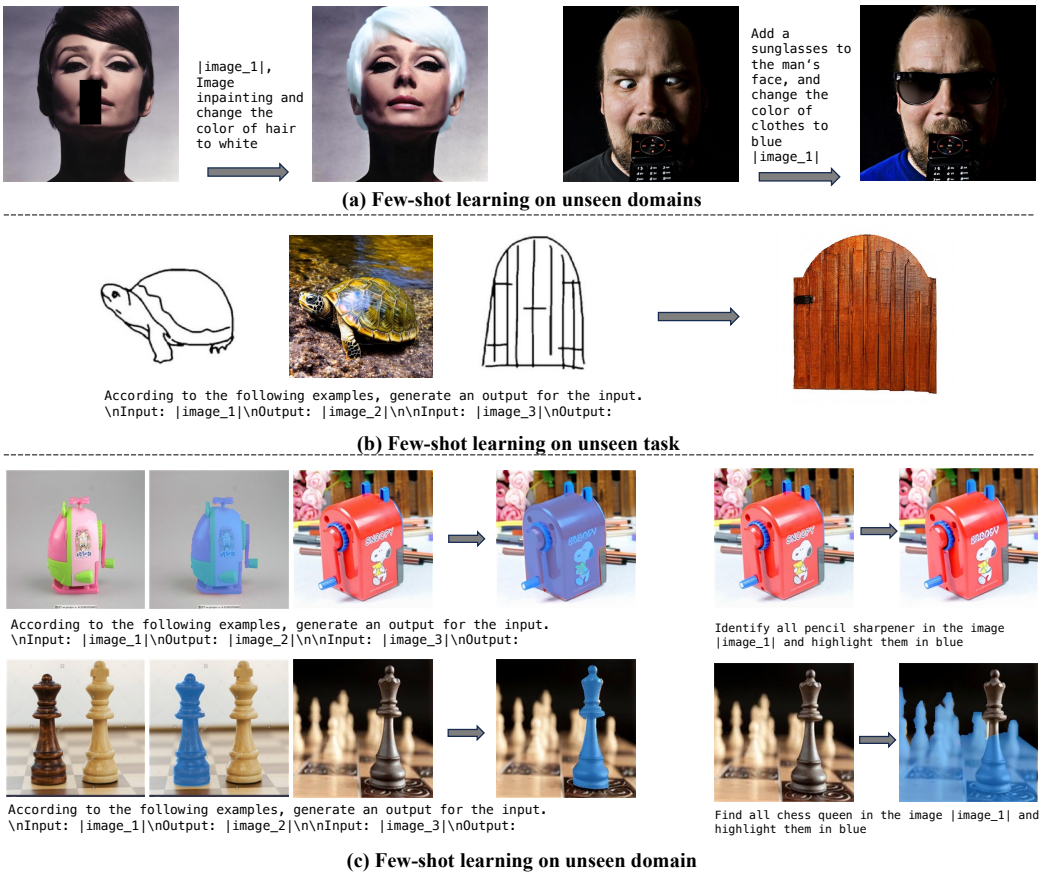


Figure 10: Examples of Emgerging Capabilities of OmniGen.

significantly simplifying the workflow and saving the cost. For instance, as shown in Figure 6, users can specify specific objects within images containing multiple elements through textual instructions ("the cat is the one in \image_2\"), and generate new images based on these instructions without needing preliminary operations such as image cropping. As shown in Figure 9, OmniGen can directly generate images based on the conditional information contained in the reference image, without loading another model to preprocess the reference image.

5.2 Reasoning Ability

We have explored the reasoning capabilities of the model and presented the results in Figure 11. As shown in the left half of Figure 11, when given an instruction without explicitly specifying the object, such as *"Where can I wash my hands? Please help me find the right place in \image_1\"*, the model can recognize image contents and infer that a sink is needed. Consequently, the model identifies and indicates the area of the sink in the image. This functionality creates potential applications in the field of embodied intelligence, assisting intelligent agents in comprehending multi-modal instructions, locating necessary objects and planning subsequent actions. Moreover, the right half of Figure 11 demonstrates that after inferring the target object, the model can also perform editing operations on it. If no object matches, the model will refrain from editing any unrelated objects.

5.3 Chain of Thought

The Chain-of-Thought (CoT) method can significantly boost the performance of LLMs by decomposing the task into multiple steps and sequentially solving each step to obtain an accurate final answer. We consider whether a similar alternative can be applied to image generation. Inspired by the basic way of human drawing, we hope to mimic the step-by-step drawing process, iteratively refining the

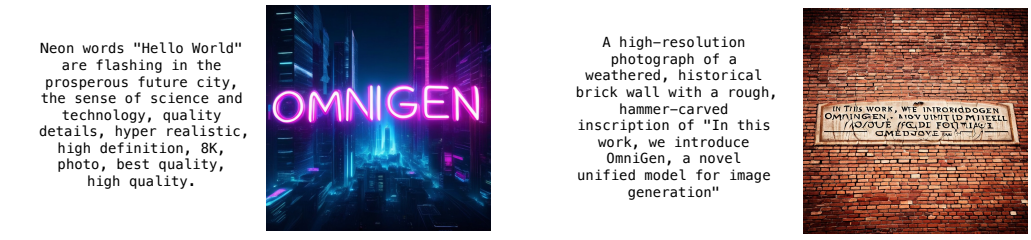
Figure 11: Reasoning Ability of OmniGen

Figure 12: Step-by-Step Image Generation Illustration Simulating the Human Drawing Process

image from a blank canvas. We construct an anime image dataset and use the PAINTS-UNDO[5] model to simulate each stage of artwork creation. We select 8 representative frames to depict the gradual development of the final image. After filtering out inconsistent sequences, we fine-tuned the model on this dataset for 16,000 steps.

The results are visualized in Figure 12, alongside the outputs generated by the original model. For step-by-step generation, the input data consists of the current step's image and text, and then the model predicts the image for the next step. It can be observed that the fine-tuned model successfully simulates the behavior of a human artist: drawing the basic outline, incrementally adding details, making careful modifications, and applying colors to the image. In this manner, users can modify the previous results to control the current output, thereby participating more actively in the image generation process, rather than passively waiting for the final image with a black-box diffusion model. Unfortunately, the quality of the final generated images does not surpass that of the original model. In the step-by-step generation approach, the model may incorporate erroneous modifications, leading to some disarray in the final image. This does not imply that the approach is unfeasible; currently, we only conduct a preliminary exploration, leaving further optimizations for future research. Based on the findings of previous work [38] on LLMs, which indicate that process supervision significantly outperforms outcome supervision, we posit that supervising the drawing process of images is a promising direction that may assist the model in handling more complex and diverse scenes.

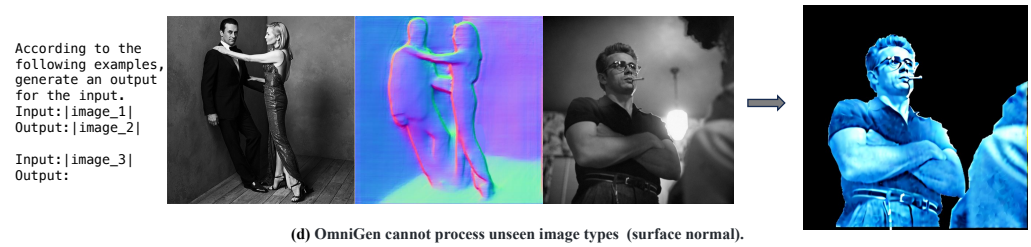
[5]https://github.com/lllyasviel/Paints-UNDO



(a) OmniGen lacks the capability to render long texts and is currently limited to processing shorter texts (as shown in the left figure).

(b) OmniGen sometimes generates incorrect details for hands.

(c) OmniGen can identify the required individual from the reference image based on instructions, and generate image maintains consistency with the reference image in terms of the person's hairstyle and forehead wrinkles; however, the overall facial details do not accurately match the original image.



(d) OmniGen cannot process unseen image types (surface normal).

Figure 13: Failure cases of OmniGen

6 Limitations and Discussions

Figure 13 illustrates several typical failure cases of the current model. We summarize the limitations of the current model as follows:

- Similar to existing diffusion models, OmniGen is sensitive to text prompts. Typically, detailed text descriptions result in higher-quality images.
- The current model's text rendering capabilities are limited; it can handle short text segments but fails to accurately generate longer texts. Additionally, due to resource constraints, the number of input images during training is limited to a maximum of three, preventing the model from handling long image sequences.
- The generated images may contain erroneous details, especially small and delicate parts. In subject-driven generation tasks, facial features occasionally do not fully align. OmniGen also sometimes generates incorrect depictions of hands.
- OmniGen cannot process unseen image types (e.g., image for surface normal estimation).

We believe that most limitations can be addressed by training the model on more related data. Moreover, compared to most models, fine-tuning OmniGen for downstream tasks is simpler, as it

inherently supports various image generation tasks without the need for extensive efforts and costs to build additional networks.

7 Related Work

7.1 Generative Foundation Models

The generative foundation model serves as the core of many contemporary artificial intelligence systems, revolutionizing the way machines interact with humans. The GPT series [54; 48] have demonstrated that language models can learn numerous tasks via training on a large-scale dataset. Following this trend, the rise of large language models (LLMs) [43; 3; 1] has further showcased their versatility, adeptly performing various tasks such as question answering, text summarization, and code generation within a single framework. Beyond language, multimodal large language models [39; 12] have been proposed to integrate vision and language capabilities. For example, as a typical architecture and popular trend, LLaVA [39] equips the LLM with visual perception and understanding capabilities by linking the vision encoder to the LLM through a connector layer. These models have shown impressive performance in vision-language understanding tasks. However, despite their ability to handle mixed text and image inputs, they lack the capability to generate images. The construction of a universal foundation model for image generation remains unclear and has not been fully explored. In this work, we propose a universal generative model that accepts arbitrary interleaved multimodal inputs and generates images, marking a significant stride towards a general-purpose image generation foundation model.

Recently, some works have explored unified models that support both text and image generation. In Chameleon [61], images and texts are both tokenized into the token sequence and modeled via discrete autoregressive modeling. Concurrent works such as TransFusion [77] and Show-O [69] unify diffusion and autoregressive methods into a single model, generating text autoregressively and images through diffusion. Nonetheless, like most existing diffusion models, they can only perform text-to-image tasks and cannot handle more complex and various visual generation tasks. The unification of tasks in visual generation remains unexplored. Unlike these efforts, our current focus is on the unification of diverse visual generation tasks. The model is capable of performing various tasks, including text-to-image generation, image editing, subject-driven generation, virtual try-on, image deblurring, human pose recognition, and more. To the best of our knowledge, this is the first model capable of unifying such a wide range of visual generation tasks. Building on this foundation, further expansion into text generation is planned as the next step in the research agenda.

7.2 Diffusion Model

Recent advancements in diffusion models have been remarkable, with notable contributions from the Stable Diffusion series [56; 52; 13], DALL-E [55], and Imagen [26]. These models are predominantly designed for text-to-image generation tasks. To facilitate visual-conditioned generation, approaches such as ControlNet [73] and T2i-Adapter [45] introduce supplementary networks integrated into existing text-to-image models, thereby enabling them to accommodate image-based conditions. StyleShot [17] incorporates a style-sensitive encoder to manipulate the style feature of the output images. InstructPix2Pix [4] addresses image editing by augmenting the model with additional input channels. SEED-X [20] and Kosmos-G [49] employ an MLLM to replace the CLIP encoder in SD, improve the performance on the specific downstream task. However, these methods are task-specific, extending the capabilities of SD by modifying the model architecture. In contrast, OmniGen is a model that natively supports various image generative tasks, unifying all tasks into a single framework. Multi-task learning enhances the model's capabilities and also leads to the emergence of new abilities. Furthermore, when addressing various real-world tasks, OmniGen no longer requires any preprocessing steps or assistance from other models.

There is some work exploring the unification of computer vision (CV) tasks [2; 65; 16; 21]. However, these efforts primarily focus on classic vision tasks and do not support general image generation tasks. Additionally, current models often underperform compared to those specifically designed and trained for corresponding tasks, limiting their practical applications in real-world scenarios. In our work, the introduction of CV tasks plays a crucial role in enabling the model to learn general knowledge, thereby enhancing its image-generation capabilities and fostering the emergence of new abilities. For instance, incorporating the human pose estimation task has led to the model's ability to generate new

images directly based on the pose of a reference image, without the need for an additional model to extract the human pose. At present, We are not obsessed with the pursuit of optimal scores on CV tasks, but leave the fine-tuning of CV task performance for future research.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[2] Roman Bachmann, Oğuzhan Fatih Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4m-21: An any-to-any vision model for tens of tasks and modalities. *arXiv preprint arXiv:2406.09406*, 2024.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. URL `https://arxiv.org/abs/2309.16609`.

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[5] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model, 2024.

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

[7] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[8] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022.

[9] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[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.

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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, 2024.

[13] 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 *Forty-first International Conference on Machine Learning*, 2024.

[14] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[16] Yulu Gan, Sungwoo Park, Alexander Schubert, Anthony Philippakis, and Ahmed M Alaa. Instructcv: Instruction-tuned text-to-image diffusion models as vision generalists. *arXiv preprint arXiv:2310.00390*, 2023.

[17] Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *arXiv preprint arXiv:2407.01414*, 2024.

[18] Peng Gao, Le Zhuo, Ziyi Lin, Chris Liu, Junsong Chen, Ruoyi Du, Enze Xie, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024.

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

[20] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation, 2024. URL `https://arxiv.org/abs/2404.14396`.

[21] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12709–12720, 2024.

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

[23] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[24] Zhen Han, Chaojie Mao, Zeyinzi Jiang, Yulin Pan, and Jingfeng Zhang. Stylebooth: Image style editing with multimodal instruction. *arXiv preprint arXiv:2404.12154*, 2024.

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

[26] Imagen-Team-Google. Imagen 3, 2024. URL `https://arxiv.org/abs/2408.07009`.

[27] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

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

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[30] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

[31] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. *arXiv preprint arXiv:2206.14180*, 2022.

[32] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024.

[33] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback, 2024. URL `https://arxiv.org/abs/2404.07987`.

[34] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2869–2878, 2020.

[35] Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, Yuyin Zhou, and Cihang Xie. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024.

[36] Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *2407.08303*, 2024.

[37] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.

[38] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[40] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[41] 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.

[42] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[43] Meta. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

[44] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[45] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, 2023. URL `https://arxiv.org/abs/2302.08453`.

[46] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017.

[47] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. *arXiv preprint arXiv:2404.19753*, 2024.

[48] OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[49] 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.

[50] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[51] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[52] 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. *arXiv preprint arXiv:2307.01952*, 2023.

[53] 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.

[54] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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

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

[57] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

[58] 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. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[59] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024.

[60] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[62] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[63] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[64] X Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024.

[65] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023.

[66] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[67] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023.

[68] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023.

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

[70] 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.

[71] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for fast image restoration and enhancement. *IEEE transactions on pattern analysis and machine intelligence*, 45(2): 1934–1948, 2022.

[72] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024.

[73] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, October 2023.

[74] 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. *Advances in Neural Information Processing Systems*, 36, 2024.

[75] Na Zheng, Xuemeng Song, Zhaozheng Chen, Linmei Hu, Da Cao, and Liqiang Nie. Virtually trying on new clothing with arbitrary poses. In *Proceedings of the 27th ACM international conference on multimedia*, pages 266–274, 2019.

[76] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

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