OmniGen: Unified Image Generation

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

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. 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. 2)* ***Simplicity****: The architecture of OmniGen is highly simplified, eliminating the need for additional plugins. Moreover, compared to existing diffusion models, it is more user-friendly and can complete complex tasks end-to-end through instructions without the need for extra intermediate steps, greatly simplifying the image generation workflow. 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 the chain-of-thought mechanism. This work represents the first attempt at a general-purpose image generation model, and we will release our resources at* https://github.com/VectorSpaceLab/OmniGen *to foster future advancements.*

1. Introduction

The pursuit of Artificial General Intelligence (AGI) has intensified the demand for generative foundation models capable of handling various 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. However, the realm of visual generation has yet to reveal a counterpart that matches the universality of LLMs. Current image generation models have demonstrated proficiency in specialized tasks. For instance, in the text-to-image generation filed, models such as the SD series [12, 51, 55], DALL-E [54], and Imagen [25] have made significant strides. Meanwhile, lots of efforts have been made to extend the capabilities of diffusion models for specific tasks, such as ControlNet [75], InstandID [64], and InstructPix2Pix [3]. Currently, for each new task, designing a specific module and fine-tuning it is necessary, which hinders the development of image generation. What's worse, these task-specific models lead to a significant issue: we cannot accomplish tasks simply through input instructions but require a complex and cumbersome workflow involving multiple models and numerous steps. For example, ControlNet needs to use a detector to extract conditions (such as pose estimation maps) and then loads the corresponding condition module. InstantID can only process single-person images and requires a face detection model to detect faces beforehand, followed by using a face encoder to encode the face. If users want to edit the generated images, additional models like InstructPix2Pix need to be loaded.

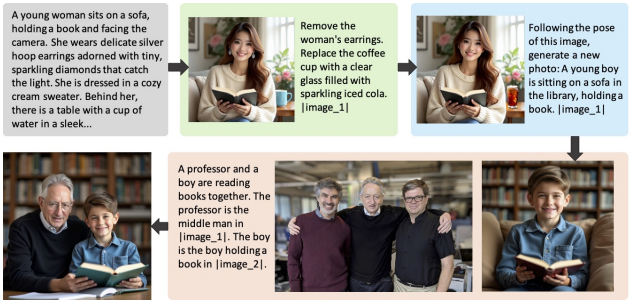

Figure 1. OmniGen can flexibly follow instructions to complete various tasks, without the need for any additional plugins. For complex tasks (e.g., examples in the second line), it can also be completed end-to-end without cumbersome intermediate steps.

Can a single model complete various tasks end-to-end through user instructions, similar to how ChatGPT handles language tasks? We envision a future where image generation is made simple and flexible: that any tasks can be

*Co-first authors
†Corresponding authors

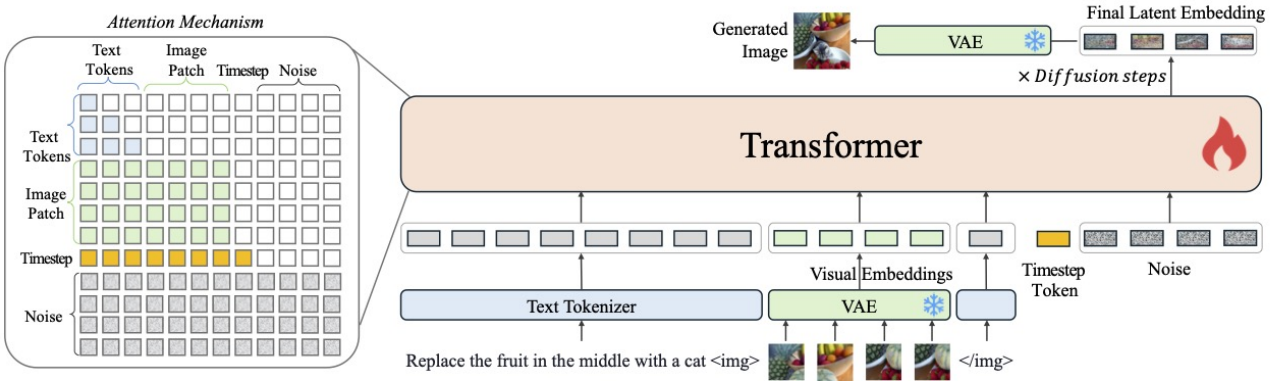

Figure 2. The framework of OmniGen. Texts are tokenized into tokens, while input images are transformed into embedding via VAE. OmniGen can accept free-form multi-modal prompts and generate images through the rectified flow approach.

accomplished directly through user instructions. Motivated by this goal, we propose a unified framework: OmniGen. As shown in Figure 1, this framework allows for convenient and flexible image generation for any purposes, where no additional plugins or operations are needed. Given the flexibility in following arbitrary instructions, the new framework also help to inspire more interesting image generation tasks.

Unlike popular diffusion models, OmniGen features a very concise structure, comprising only two main components: a VAE and a transformer model. OmniGen supports arbitrarily interleaved text and image inputs as conditions to guide image generation, instead of only accepting pure text or image. To train this architecture as a unified model, we construct a large-scale multi-task image generation dataset X2I, which unifies different tasks with one uniform format. We evaluate the well-trained model based on multiple benchmarks, demonstrating its superior generation capability compared to existing models. Remarkably, OmniGen enables effective knowledge transfer across different scenarios, allowing it to handle unseen tasks and domains while also fostering the emergence of new abilities.

Our contributions are summarized as follows:

- We introduce OmniGen, a unified image generation model that excels in multiple domains. The model demonstrates competitive text-to-image generation capabilities and inherently supports a variety of downstream tasks, such as controllable image generation and classic computer vision tasks. Furthermore, it can handle complex tasks end-to-end without any lengthy intermediate steps. To our knowledge, OmniGen is the first image generation model to achieve such comprehensive functionality.
- We construct the first comprehensive image generation dataset named X2I, which stands for "anything to image". This dataset covers a wide range of image generation tasks, all of which are 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, we explore the model's reasoning capabilities and chain-of-thought mechanism.
- OmniGen takes an initial step toward a foundational model for general image generation. We will open-source the relevant resources (model, code, and data) and hope this work can provide some insights for future image generation models.

2. OmniGen

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

2.1. Model Design

Network Architecture. The design principles of OmniGen are universality and conciseness. As illustrated in Figure 2, OmniGen adopts an architecture comprised of a Variational Autoencoder (VAE) [27] and a pre-trained large transformer model. Specifically, we use the VAE from SDXL [51] to extract continuous visual features from images. The transformer model initialized by Phi-3 [1] generates images based on instructions that serve as conditions. Only VAE is frozen during training. 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 inputs 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 [8, 64, 68, 70, 72] which lacks interaction between different modality conditions.

Input Representation. The input to the model can be multi-modal interleaved text and images in free form. We utilize the tokenizer of Phi-3 to process text. For images, we firstly employ a VAE with a simple linear layer to extract latent representations. Then, they are flattened into a sequence of visual tokens by linearly embedding each patch in the latent space. Following [49], the patch size is set to 2.

We apply standard frequency-based positional embeddings to visual tokens, and process images with varying aspect ratios as SD3 [12]. In addition, we encapsulate each image sequence with two special tokens: "<img>" and "</img>" before inserting it into the sequence of text tokens. We also append the timestep embedding [49] at the end of the input sequence. We do not need any task-specific special tokens to indicate the task type.

Attention Mechanism. Different from text, which can be decomposed into discrete tokens to model, we argue that the image should be modeled as a whole based on its nature. Therefore, we modify the common causal attention mechanism in LLM by 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 decoder to decode the latent representation into the predicted image. 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 [39] to optimize the parameters of model. Different from the DDPM [24], 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:

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

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 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 significantly higher weights than those without changes, guiding the model to focus on the areas to be modified.

Stage	Resolution	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. Details about each training stage of OmniGen.

Training Pipeline. Following previous work [5, 12, 16], 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 1. We adopt the AdamW [40] with $\beta = (0.9, 0.999)$ as the optimizer. All experiments are conducted on 104 A800 GPUs.

3. X2I Dataset

To achieve robust multi-task processing capabilities, it is essential to train the model 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 all data into a unified format, and Figure 3 presents some examples of the X2I dataset. The entire dataset comprises approximately 0.1 billion images. A detailed description of the composition will be provided in the following sections.

3.1. Text to Image

The input for this subset of data is plain text. We collect multiple open-source datasets from various sources: Recap-DataComp [33](a subset of 56M images), SAM-LLaVA [5], ShareGPT4V [6], LAION-Aesthetic [57](a subset of 4M images), ALLaVA-4V [4], DOCCI [45], DenseFusion [34] and JourneyDB [59]. 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 an internal collection of 16M high-quality images to enhance the aesthetic quality of the generated images. We use the InternVL2 [10] to create synthetic annotations for internal data and LAION-Aesthetic.

Figure 3. Examples of X2I dataset. We standardized the input of all tasks into an interleaved image-text sequence format. The placeholder |image_i| represents the position of the i-th input image.

3.2. Multi-modal to Image

Different from existing diffusion models, OmniGen can accept more flexible multi-modal instruction, thus adapting to a variety of data and task types.

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 (MagicBrush [74], InstructPix2Pix [3] and SEED-edit [17],), human motion (Something-Something [21]), virtual try-on (HR-VITON [30] and FashionTryon [77]), and style transfer (StyleBooth [23]). The issue of utilizing additional visual conditions for finer-grained spatial control has garnered widespread attention[31, 75]. We employ the MultiGen [52] dataset to learn this function, and select six representative visual conditions: Canny, HED, Depth, Skeleton, Bounding Box, and Segmentation. These types of tasks take text prompts with specific visual conditions (such as segmentation maps, and human pose maps) as multi-modal inputs. We standardize all tasks into the input-output pair format as shown in Figure 3-(b).

3.2.2. Subject-driven Image Generation

We construct 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 leverage the GRIT dataset [50], which annotates object names within images. Using these annotations, we apply GroundingDINO [38] for open-set text-to-bounding-box detection. Based on the grounded bounding boxes, we employ SAM [28] to segment object masks from cropped images. We further use the MS-Diffusion [66] to repaint the object images for higher quality and diversity of data. The process of data construction and the final instruction format are illustrated in supplementary material. Through these methods, we acquire 6M data pairs.

Although the GRIT-based approach provides a substantial amount of samples, the input data extracted directly from original images will tend to lead the model to fall into a simple copy-and-paste pattern. To fully unleash the subject-driven image generation capability of OmniGen, we construct a high-quality natural images dataset from the web. We use the search engine to collect multiple different images of the same person, using one as the input and another as the output. An example is shown in Figure 3-(c). More details about the construction process are given in the supplementary material.

3.2.3. Computer Vision Tasks

We introduce classic computer vision tasks to expand the boundaries of the model's generative capabilities. For low-level vision tasks (low-light image enhancement [67], deraining [73], deblurring [44], inpainting [52], outpainting [52] and colorization [57]), where the annotation itself is an image, we only add text instructions, which are randomly sampled from instructions generated by GPT-4o. For high-level vision tasks, we choose to represent all annotations as images. We use images from LAION [57] as source images and annotations from [52] as targets to construct image pairs (such as source image and its human pose mapping). The annotations include human pose, depth mapping, canny, Hed edge, and segmentation. Additionally, we also use several datasets for referring image segmentation, including RefCOCO [26], ADE20k [78], and ReasonSeg [29]. As shown in Figure 3-(d), the input is the source image and a natural language expression, the output is an image with the corresponding object highlighted in blue.

3.3. Few-shot to Image

We build a few-shot to image dataset to stimulate the in-context learning ability of the model. 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 is shown in Figure 3-(e). Due to limitations in training

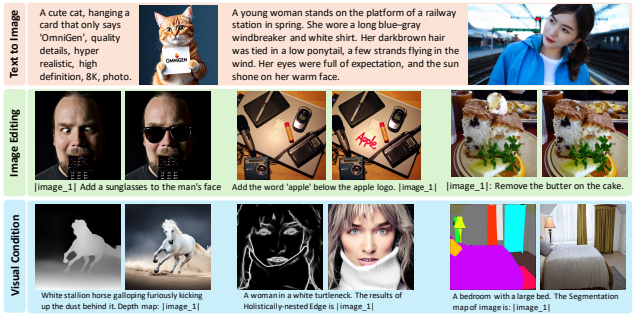

Figure 4. The results of different image generation tasks.

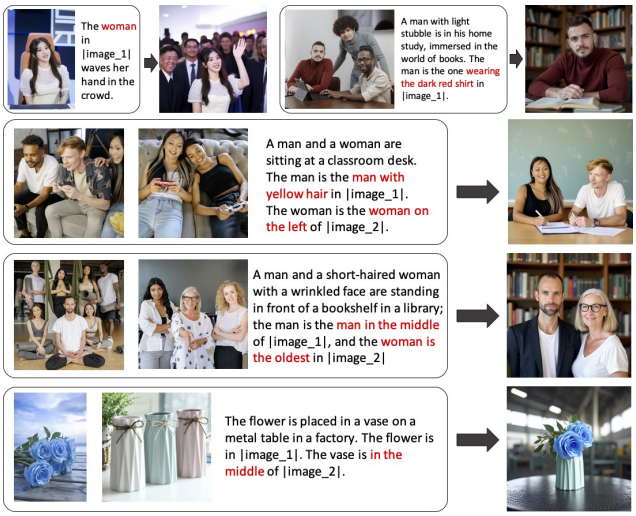

Figure 5. Subject-driven generation. When the reference image contains multiple objects, OmniGen can automatically identify the required objects based on textual instructions without the need for additional preprocessing steps like image cropping or face recognition.

resources, we opted to use only one example to improve training efficiency.

4. Experimental Results

The default inference step is set to 50. We use the classifier-free guidance for both text condition and image condition following [3]. The default guidance scale is set to 2.5, and the image guidance scale is set to 1.6. For image editing task, we increase the image guidance scale to 2.

4.1. Image Generation

4.1.1. Qualitative Results

Figure 4 summarizes the results of different image generation tasks, including text-to-image, image editing, and visual conditional. These results demonstrate that OmniGen can handle various downstream tasks based on multi-modal instructions.

Figure 5 presents the outcomes of the subject-driven generation task. OmniGen 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 one wearing the dark red shirt in image) without requiring additional preprocessing steps. In contrast, existing models [22, 64] require manually cropping the target person from a multi-person image, detecting faces with a face detector, encoding faces with a face encoder, and inputting them into the diffusion model through a cross-attention module. Without these complex processes, OmniGen can complete the task in a single end-to-end step.

More qualitative results are provided in supplementary materials.

Model	Params	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*	0.61	0.97	0.74	**0.66**	0.81	0.13	0.35
SD3	8.0B+4.7B*	0.68	0.98	0.84	**0.66**	0.74	0.40	0.43
OmniGen	3.8B	**0.70**	**0.99**	0.86	0.64	**0.85**	0.31	**0.55**

Table 2. Results of GenEval Benchmark. * means the parameter of the frozen text encoder. Our model achieves comparable performance with a relatively small parameter scale.

4.1.2. Text to Image

Following [12], we evaluate text-to-image generation capability on the GenEval [20] benchmark. We compare the performance of OmniGen with the reported results of other popular image generation models, as summarized in Table 2. Surprisingly, OmniGen achieves similar performance compared to the current diffusion models, such as SD3, which underscores the effectiveness of our framework. Notably, OmniGen has only 3.8 billion parameters, whereas the SD3 model has a total of 12.7 billion parameters. The architecture of OmniGen is significantly simplified, eliminating the cost of an additional encoder, thereby greatly enhancing the efficiency of parameter utilization.

4.1.3. Multi-modal to Image

CLIP Simi	-T↑	-I↑
I-Pix2Pix [3]	0.219	0.834
MagicBrush [74]	0.222	0.838
PnP [63]	0.089	0.521
Null-Text [42]	**0.236**	0.761
EMU-Edit [58]	0.231	**0.859**
OmniGen	0.231	0.829

CLIP Simi	-T↑	-I↑
Tex-Inv [13]	0.255	0.780
DreamBooth [56]	0.305	0.803
ReImagen [7]	0.270	0.740
SuTI [9]	0.304	0.819
Kosmos-G [47]	0.287	**0.847**
OmniGen	**0.315**	0.801

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

We evaluate the image editing on EMU-Edit [58] dataset and subject-driven generation capability on DreamBench [56]. We use CLIP-T to measure how well the model followed the instructions, while CLIP-I similarity scores

	Seg. Mask (mIoU↑)	Canny Edge (F1 Score↑)	Hed Edge (SSIM↑)	Depth Map (RMSE↓)
T2I-Adapter [43]	12.61	23.65	-	48.40
Gligen [35]	23.78	26.94	0.5634	38.83
Uni-ControlNet [76]	19.39	27.32	0.6910	40.65
UniControl [52]	25.44	30.82	0.7969	39.18
ControlNet [75]	32.55	34.65	0.7621	35.90
ControlNet++ [31]	**43.64**	37.04	0.8097	**28.32**
OmniGen	40.06	**38.96**	**0.8332**	31.71

Table 4. Comparison with SOTA methods on controllable image generation. ↑ indicates higher result is better, while ↓ means lower is better.

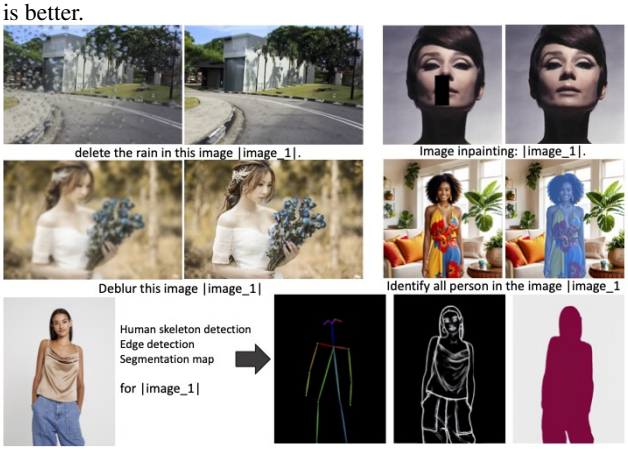

Figure 6. The results of OmniGen in various traditional vision tasks.

measure the model's ability to preserve elements from the source image. As shown in Table 3, OmniGen exhibits comparable performance to the current state-of-the-art models. Noted for subject-driven generation task, we select one image of the specific object as input instead of fine-tuning the model like DreamBooth.

In Table 4, we use the dataset and script from [31] to evaluate the generation capability based on visual conditions. 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. We can find that OmniGen achieves competitive results for all conditions.

4.2. Computer Vision Tasks

We present several qualitative results of computer vision tasks in Figure 6. OmniGen can handle various low-level vision tasks such as deraining, deblurring, and inpainting. At the bottom of Figure 6, we can see that OmniGen is also able to handle high-level vision tasks, such as human pose recognition.

Note that OmniGen's capabilities in CV tasks are not intended to surpass existing state-of-the-art models that developed for specific tasks over a long period. We hope it can transfer the knowledge gained from these traditional computer vision tasks to image generation tasks, thereby unlocking greater potential. Some examples are shown in

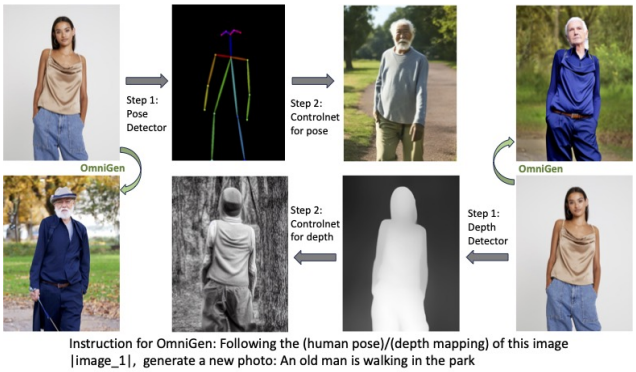

Figure 7. ControlNet involves multiple steps and plugins: firstly using the corresponding detector to extract information from the reference image, and then loading the appropriate ControlNet module to process the visual conditions. In contrast, OmniGen completes the entire task in a single step within a single model.

Figure 7. 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. Can we directly use OmniGen to generate new images based on a reference image in only one step? Surprisingly, even without having encountered such a task before, OmniGen handles it admirably without explicit extraction of additional conditional images. More specifically, we can directly input the reference image and text instruction (e.g., *Follow the human pose of this image to generate new image.*) to generate the target image in only one step without any explicit additional intermediate procedures.

5. Further Analysis

LLMs demonstrate remarkable generalization capabilities, and 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.

Task Composition. In real-world applications, user requirements often involve combinations of tasks. As shown in Figure 8-(a), our model is capable of simultaneously processing multiple instructions, including those for different tasks as well as multiple instructions for the same task. These results highlight our model's versatility and potential for widespread adoption in the wild.

End-to-end Workflow. Users typically need to load multiple models and perform multi-step processing to ultimately

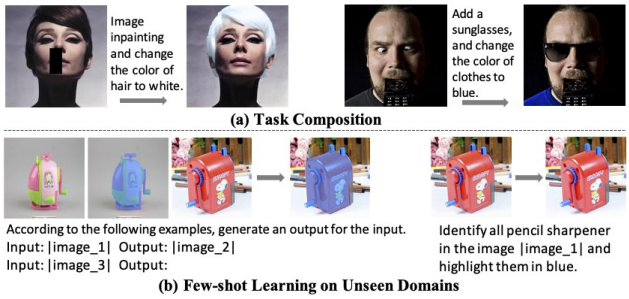

Figure 8. Examples of Emgerging Capabilities of OmniGen.

Figure 9. Reasoning ability of OmniGen.

Figure 10. Simulate the process of human drawing via step-by-step generation.

generate a satisfactory image, making the workflow very cumbersome and costly. OmniGen possesses both excellent multi-modal understanding and image generation capabilities, enabling it to complete a lot of tasks without relying on external models, thereby significantly simplifying the workflow and saving cost. As demonstrated in Figure 7, OmniGen can extract the relevant conditional information from the reference image and generate a new image based on the captured condition within one step. This process is implicit, with all processing completed internally within the model, only requiring the user to input a simple command.

In-context Learning for Unseen Domains. We use the data from FSS [32] which contains objects that have never been seen in previous datasets to evaluate the generalized ability. In the right of Figure 8-(b), we can see that OmniGen is not familiar with the concepts of "pencil sharpeners", and it cannot identify it from images. However, when provided with an example, the model is capable of making accurate predictions. By providing an example, the model is capable of successfully completing this unseen segmentation task.

5.2. Reasoning Ability

We have explored the reasoning capabilities of the model and presented the results in Figure 9. As shown in the left half of Figure 9, 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 9 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(see the example in the bottom right corner of Figure 9).

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 image from a blank canvas. We construct an anime image dataset and use the PAINTS-UNDO[1] 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 10, 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 [36] on LLMs, which indicate that process supervision significantly outperforms outcome supervision,

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

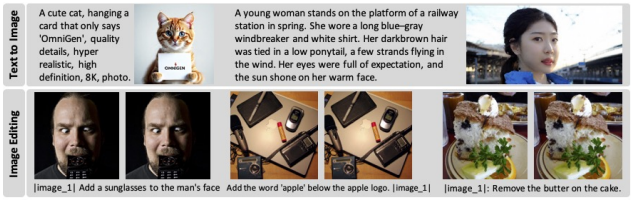

Figure 11. **Top**: results of causal attention; **Bottom**: results without loss weights.

Input Image Representation	EMU-Edit		DreamBench	
	CLIP-T	CLIP-I	CLIP-T	CLIP-I
CLIP	0.232	0.820	0.312	0.789
VAE	0.231	0.829	0.315	0.801

Table 5. Comparison of different methods to obtain visual embedding for input images.

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.

6. Ablation Study

In this section, we assess the necessity of certain modules in OmniGen

Attention Mask. Figure 11-Top shows the generated images without the modified attention in Section 2.1. As we can see, there is a lot of noise in the images. Compared to Figure 4, the generated images in Figure 11-Top have many distorted areas, confirming that the unidirectional attention mechanism in LLMs is not suitable for the rectified flow.

Weighted Loss. Figure 11-Bottom shows the results when the weighted loss from Section 2.2 is not used for training. For image editing tasks, since the edited area is usually small, the model tends to learn to copy the input image directly as the output. Comparing Figure 11-Bottom and Figure 4, it's evident that using weighted loss can prevent the model from learning this shortcut.

Input Image Representation. The CLIP model is widely used in multi-modal understanding models [37]. We also explored the difference between using CLIP[2] and using VAE to process input images. The results are shown in Table 5. There is not much difference between the two strategies. Using VAE has a slight advantage in image similarity, which might be related to the fact that the output images also require VAE processing. Another advantage of using VAE is that it avoids additional modules, maintaining the simplicity of the model. Directly inputting the original image is another simpler approach that we leave for future exploration.

7. Related Work

Generative Foundation Models. The GPT series [46, 53] have demonstrated that language models can learn numerous tasks via training on a large-scale dataset. Beyond language, multi-modal large language models [11, 37] have been proposed to integrate vision and language capabilities. However, they lack the capability to generate images. Some works propose integrating LLMs with diffusion models to equip LLMs with image generation capability [18, 60, 61, 69]. Others use discrete tokens to support both image and text generation simultaneously [41, 62, 71]. They focus on multi-modal generation with limited image-generation capability. Concurrent works such as TransFusion [79] and Show-O [71] unify diffusion and autoregressive methods into a single model, generating text autoregressively and images through diffusion. Nonetheless, like existing diffusion models, these works focus on a limited range of image generation tasks, primarily text-to-image generation, and cannot cover more complex and various visual generation tasks. The construction of a universal foundation model for image generation remains unclear and has not been fully explored.

Diffusion Model. Recent advancements in diffusion models have been remarkable, with notable contributions from the Stable Diffusion (SD) series [12, 51, 55], DALL-E [54], and Imagen [25]. These models are predominantly designed for text-to-image generation tasks. Many efforts have been made to extend the capabilities of diffusion models, such as ControlNet [75], T2I-Adapter [43], StyleShot [15]. InstructPix2Pix [3], InstructSD [48] and EMU-edit [58] explore performing general image editing tasks through instructions. 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-generation tasks, and no longer requires any preprocessing steps or assistance from other models. There is some work exploring the unification of computer vision (CV) tasks [2, 14, 19, 65]. However, these efforts primarily focus on classic vision tasks instead of general image generation tasks, and often underperform compared to those specifically designed and trained for corresponding tasks.

8. Limitations and Conclusion

In this work, we introduce OmniGen, the first unified image generation model. It is designed to be simple, flexible, and easy to use. We believe that the future paradigm of image generation should be simple and flexible, allowing for the direct generation of various images through any multimodal instructions without complex workflows.

However, the current model still has some issues. The text rendering capability is limited. The output images may contain undesired details (e.g. abnormal hand). Besides, previously unseen image types (e.g., surface normal map) can hardly be processed as expected. More failure cases are provided in supplementary materials.

[2]https://huggingface.co/openai/clip-vit-large-patch14-336

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

[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. 8

[3] 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. 1, 4, 5, 8

[4] 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. 3

[5] 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. 3

[6] 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. 3

[7] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 5

[8] 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. 2

[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. 5

[10] 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. 3

[11] 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. 8

[12] 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. 1, 3, 5, 8

[13] 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. 5

[14] 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. 8

[15] 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. 8

[16] 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. 3

[17] 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. 4

[18] 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. *arXiv preprint arXiv:2404.14396*, 2024. 8

[19] 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. 8

[20] 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. 5

[21] 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. 4

[22] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. *arXiv preprint arXiv:2404.16022*, 2024. 5

[23] Zhen Han, Chaojie Mao, Zeyinzi Jiang, Yulin Pan, and Jingfeng Zhang. Stylebooth: Image style editing with multimodal instruction. *arXiv preprint arXiv:2404.12154*, 2024. 4

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[25] Imagen-Team-Google. Imagen 3, 2024. 1, 8

[26] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in pho-

tographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 4

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

[28] 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. 4

[29] 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. 4

[30] 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. 4

[31] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback, 2024. 4, 6

[32] 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. 7

[33] 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. 3

[34] 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. 3

[35] 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. 6

[36] 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. 7

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 8

[38] 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. 4

[39] 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. 3

[40] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[41] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024. 8

[42] 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. 5

[43] 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. 6, 8

[44] 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. 4

[45] 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. 3

[46] OpenAI. Gpt-4 technical report, 2024. 8

[47] 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. 5

[48] Sayak Paul. Instruction-tuning stable diffusion with instructpix2pix. *Hugging Face Blog*, 2023. https://huggingface.co/blog/instruction-tuning-sd. 8

[49] 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. 2, 3

[50] 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. 4

[51] 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. 1, 2, 8

[52] 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. 4, 6

[53] 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. 8

[54] 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. 1, 8

[55] 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. 1, 8

[56] 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. 5

[57] 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. 3, 4

[58] 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. 5, 8

[59] 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. 3

[60] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 8

[61] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context interleaved and interactive any-to-any generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27425–27434, 2024. 8

[62] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 8

[63] 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. 5

[64] 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. 1, 2, 5

[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. 8

[66] 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. 4

[67] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 4

[68] 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. 2

[69] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal LLM. In *Proceedings of the International Conference on Machine Learning*, pages 53366–53397, 2024. 8

[70] 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. 2

[71] 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. 8

[72] 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

[73] 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. 4

[74] 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. 4, 5

[75] 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, 2023. 1, 4, 6, 8

[76] 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. 6

[77] 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. 4

[78] 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. 4

[79] 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. 8