VisualCloze: A Universal Image Generation Framework via Visual In-Context Learning

Zhong-Yu Li[1,4,5*] Ruoyi Du[3,5*] Juncheng Yan[5,6] Le Zhuo[5]
Zhen Li[7†] Peng Gao[2,5] Zhanyu Ma[3] Ming-Ming Cheng[1,4†]
[1]NKIARI, Shenzhen Futian
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3]Beijing University of Posts and Telecommunications [4]VCIP, CS, Nankai University
[5]Shanghai AI Laboratory [6]Tsinghua University [7]The Chinese University of Hong Kong

Abstract

Recent advances in diffusion models have significantly advanced image generation; however, existing models remain task-specific, limiting their efficiency and generalizability. While universal models attempt to address these limitations, they face critical challenges, including generalizable instruction design, appropriate task distributions, and unified architectural design. In this work, we propose VisualCloze, a universal image generation framework, to tackle these challenges. Unlike existing methods that rely on language-based task descriptions, leading to task ambiguity and weak generalization, we integrate visual in-context learning, allowing models to identify tasks from demonstrations. Meanwhile, the inherent sparsity of visual task distributions hampers the learning of transferable knowledge across tasks. To this end, we introduce Graph200K, a graph-structured dataset that establishes various interrelated tasks, enhancing task density and knowledge transfer. Furthermore, we uncover an intrinsic alignment between image infilling and in-context learning, enabling us to leverage the strong generative priors of pre-trained infilling models without modifying their architectures. Experiments demonstrate that VisualCloze achieves strong performance across various in-domain tasks while generalizing to unseen tasks in few-shot and zero-shot settings. Our codes and dataset are available at https://visualcloze.github.io/.

1. Introduction

Recent advancements in image generation, propelled by the progress of diffusion models [14, 30, 78], have led to a wide range of applications, including image editing [60], style transfer [55, 71], virtual try-on [10, 11], and personalized generation [34, 48], among others. However, these tasks typically require task-specific models, which limit efficiency and scalability for real-world applications. In recent years, there has been growing interest in unified generative models [26, 35, 40], which aim to handle diverse image generation tasks within a single framework. However, key challenges remain despite significant progress, *e.g.*, instruction design, task distribution, and model architecture.

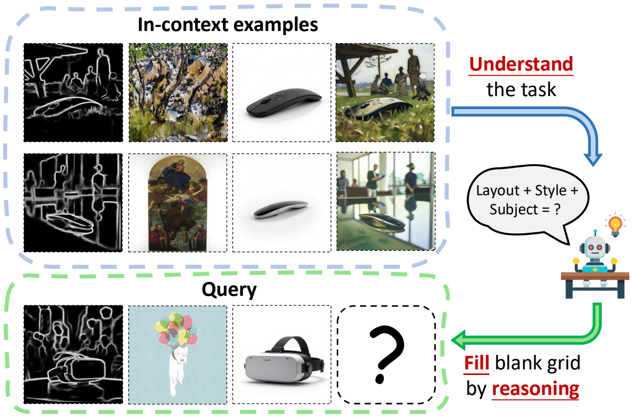


Figure 1. Illustration of our universal image generation framework based on visual in-context learning. Given one query of a specific task, the generative model learns the task by observing a few in-context examples presented as demonstrations.

Regarding instructions that guide the model to differentiate distinct tasks and produce expected behavior, existing methods primarily rely on language instructions [26, 40] or task-specific tokens [35]. However, the task complexity and inherent gap between vision and language modalities make it hard for the model to understand the task intent based on language description alone. This issue leads to task confusion [35] and hinders generalization on unseen tasks [32, 61]. In contrast, large language models have

[*]Equal contribution.
[†]Corresponding author.

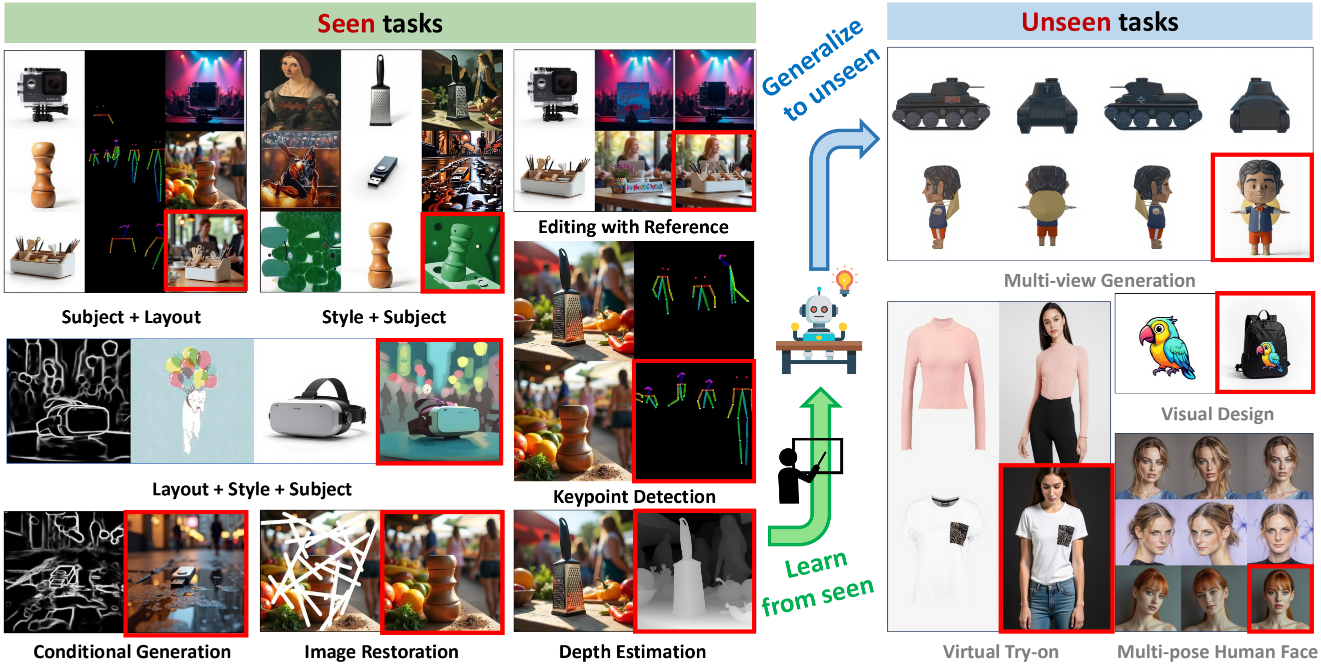


Figure 2. Results of our universal image generation framework. The grid image follows the structure of Fig. 1, and the red box indicates the generated image. After training on a wide range of tasks, our framework exhibits a certain generalization ability on unseen tasks.

successfully achieved unified multi-task modeling, partially due to the rise of in-context learning [5], which allows models to understand tasks using only a few demonstrations. In the vision modality, prior works [1, 3, 4, 39, 57, 61, 72] also show the potential of in-context learning, even though they are still limited to limited task domain (e.g., dense prediction, style transfer, *etc.*) [58, 77] or simple generation tasks that involve only one condition and one target image [39, 52]. Inspired by this, not only language instructions, we also adopt a few in-context task examples as visual instructions to guide the model toward desired behavior, as shown in Fig. 1. Then, the model supports various tasks with mitigated task ambiguity and supports few-shot generalization on unseen tasks, as shown in Fig. 2. The model also exhibits a certain degree of zero-shot generalization.

From the perspective of task distribution, visual tasks are inherently sparse compared to those in natural language processing because task-specific datasets [61, 75] for different tasks have minimal overlap [18, 29, 69]. Such sparse task learning isolates the knowledge of each task and limits the model from learning shared features across tasks. Moreover, the weak correlations between tasks hinder knowledge transfer and adaptability of models. However, existing works in multi-task learning [9, 15, 28, 47] have verified the benefits of overlapping knowledge across related tasks. To alleviate the sparsity of visual tasks, we introduce a graph-structured dataset, Graph200K, where each image is associated with annotations spanning five meta-tasks, *i.e.*, conditional generation [70], IP preservation [66], style transfer [71], image editing [60], and restoration [67]. By combining different conditions, we train the model with a variety of tasks that overlap and interact with each other. Given this highly overlapping task space, our dataset significantly increases task density, allowing the model to learn shared and transferable representations more effectively.

For the architecture design, to fully realize the potential of universal image generation, it is crucial to design structures that accommodate flexible task formats [26, 32, 61], especially when using in-context examples, and are also compatible with state-of-the-art models [30, 78] to leverage their strong generative priors. In this work, we find that the state-of-the-art image infilling model [30] has a consistent objective with our in-context learning based universal generative models. Specifically, we concatenate all input and output images as a grid-layout image, where the objective of a task is to fill the unknown grid, as illustrated in Fig. 1. This alignment enables us to build our model upon advanced general-purpose infilling models without additional modifications, thereby achieving powerful universal generation capabilities with minimal data and training costs.

In this work, we propose a universal image generation framework, VisualCloze, which fine-tunes FLUX.1-Fill-dev [30] with interrelated tasks sampled from Graph200K to learn transferable knowledge and support visual in-context learning. As the number of in-context examples increases, we observe enhanced performances and reduced task confusion, enabling the model to support a broad spectrum of in-domain tasks, including conditional generation, image restoration, editing, style transfer, IP-preservation, and their combinations. On unseen tasks, the model also shows a certain degree of generalization ability, as shown in Fig. 2. In summary, our main contributions are as follows:

- We propose an in-context learning based universal image generation framework that supports a wide range of in-domain tasks and exhibits generalization to unseen ones.
- We design a graph-structured dataset, Graph200K, which constructs a compact task space, enabling flexible online task sampling and promoting the models to learn shared and transferable representations across tasks.
- We find that the state-of-the-art infilling model shares a consistent objective with unified image generation, enabling us to achieve exceptional performance through minimal tuning without modifying the model structure.

2. Related Work

2.1. Image Generation

Recent advances in text-to-image generation have achieved remarkable performance, largely driven by the development of autoregressive models [37, 50, 68] and diffusion models [2, 12, 14, 17, 23, 36, 38, 43, 46]. Among these, rectified flow transformers [14, 16, 30, 78] have shown great training efficiency and overall performance. Building on these foundational models, diverse applications have emerged, such as conditional generation [70], style transfer [55], and personalized generation [34]. More recently, universal models that address various tasks [32, 40, 73] have been explored. For example, unified models like OmniGen [61] leverage large vision language models to consolidate multiple tasks into a single framework. Similarly, UniReal [8] unifies image generation tasks as discontinuous video generation. However, they still face issues such as over-reliance on language instructions, isolation and sparsity of visual tasks, and architecture design that accommodates flexible task formats. To address these issues, we propose a universal image generation framework that unifies generation tasks as image infilling. Through visual in-context learning and our Graph200K dataset that constructs a denser task space to learn transferable representations, our method alleviates ambiguity to support a diverse set of in-domain tasks and generalizes to tasks unseen during training.

2.2. Visual In-context Learning

Along with the emergence of large language models, such as GPT-3 [5], in-context learning [13] has been an effective approach to allow the language model to understand and perform complex tasks given a few demonstrations. Early works [20, 21] in vision modality propose image analogies to automatically create an image filter from examples. In recent years, leveraging inpainting model [3, 4, 72], masked image modeling [39, 57, 58], or vision-language model [1, 76], visual in-context learning is proposed to handle more tasks. However, they mainly focus on dense prediction [49, 51, 77] or visual understanding [54]. OmniGen [61] also leverages in-context learning to generalize to unseen domains, *e.g.*, segmenting unseen concepts when the model has learned the segmentation task during training. However, it mainly focuses on simple tasks of dense prediction, and the gap between the unseen and training domains is still limited. Some recent works [31, 39, 52, 59] extend visual in-context learning to image generation, but they are still limited by simple tasks such as conditional generation and dense prediction. Moreover, the sparsity of visual tasks makes it difficult for models to learn transferable and overlapping knowledge across tasks, limiting the generation ability of in-context learning. In contrast, we introduce a graph-structured dataset that supports interrelated tasks and thus constructs a more dense task space, promoting the model to learn shared and transferable knowledge and enhance its adaptability.

3. Dataset

Recent works [25, 61] have made great progress in unified image generation. However, their ability to generalize to unseen tasks remains highly limited. We partially attribute this to the sparsity and isolation of visual tasks, making it difficult for the model to capture shared features across tasks and perform well on unseen ones. Moreover, weak correlations between tasks further hinder knowledge transfer, restricting the adaptability of models. Therefore, increasing task density or strengthening task inter-relations helps improve the generalization ability of models via a compact task distribution. In this paper, taking the Subject200K [53] dataset as a testbed, we extend each data point in this dataset into a data graph, constructing an expanded dataset dubbed Graph200K. As illustrated in Fig. 3 (a), we can flexibly construct various related tasks by sampling different paths within the data graph.

3.1. Graph-Structured Multi-Task Dataset

In natural language processing, tasks overlap significantly, facilitating strong cross-task learning ability. In contrast, visual tasks are inherently distinct, posing challenges for vision models to achieve similar generalization ability via instruction tuning. To ease this issue, we introduce a *Graph-Structured Multi-Task Dataset*. As illustrated in Fig. 3 (a), given a text-to-image dataset, each original image O_n is treated as the central node of a graph G_n around which diverse task annotations are constructed, including those for various spatial conditions (C_n^i), various degradations (D_n^i), various image editing results (E_n^i), reference image (R_n) for IP-preservation, and style transfer (T_n^i) with various reference styles (S_n^i). The construction process for each task pair is detailed in the next section.

As shown in Fig. 3 (a), each task annotation forms a bidirectional edge with the image. Thus, the graph is strongly

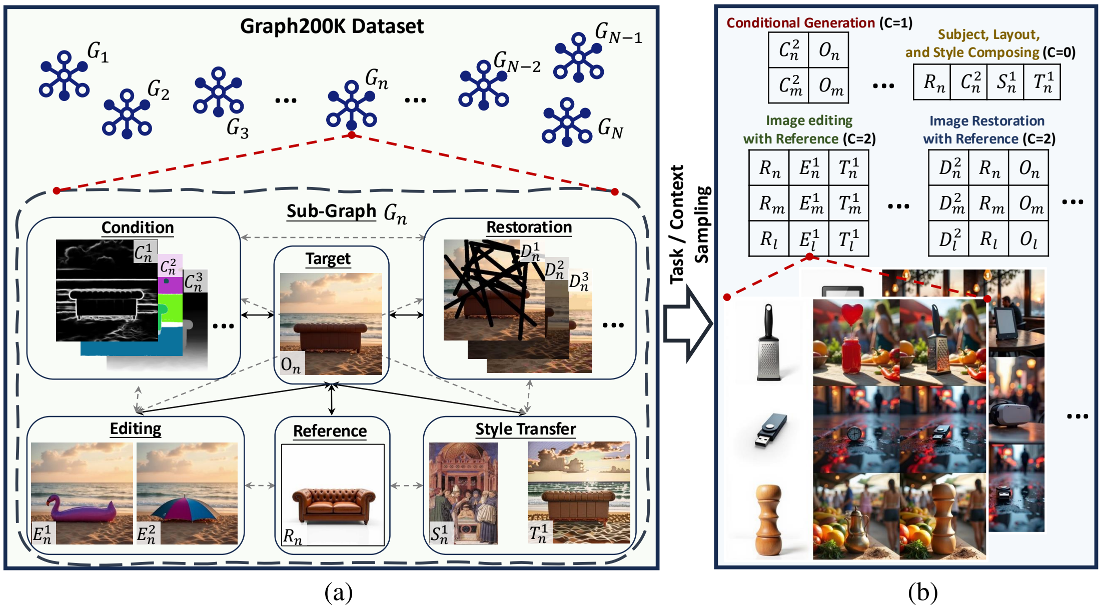


Figure 3. (a) Graph200K consists of N graph, defined as $\{G_1, \ldots, G_n, \ldots, G_N\}$, where the n-th graph contains an image (O_n) and its annotations of various tasks, such as conditional generation (C_n^i), restoration (D_n^i), editing (E_n^i), IP preservation (R_n^i), and style transfer (T_n^i). The superscript i indicates the fine-grained classification of tasks. (b) Through sampling nodes along the graph, Graph200K supports diverse and flexible tasks. For a sample with $R + 1$ rows, the first R rows mean in-context examples defined in Section 4.1.

connected, which means that for any two nodes, bidirectional paths exist between them. In other words, a generation task can be formulated as a path within the graph. The nodes along these paths (except the end nodes) serve as condition images, which is analogous to the question in instruction fine-tuning, while the target image (the end nodes) plays the role of the answer. Specifically, there are 49 types of nodes in our Graph200K and we sample up to 134 highly overlapping tasks, making the model learn more compact and shared representations across tasks. Moreover, it enriches the diversity and flexibility of our instruction fine-tuning data. For example, the path $\{R_n, E_n, O_n\}$ from graph G_n corresponds to the image editing with subject reference, as shown in Fig. 3 (b) bottom.

3.2. Dataset Construction

For convenience, we inherit subject-driven data from the Subjects200K [53]. Additionally, 32 different degradations are applied online to the images to acquire restoration data. We summarize the data construction methods in this section for the remaining three tasks.

Conditional generation. Each image is paired with 12 distinct conditions generated by specialized models, including canny edges [6], HED edges [62], Hough lines [19], semantic segmentation maps [33], depth maps [64], shape normal maps [63], and human keypoints [7], following ControlNet [70]. This work extends the conditions by incorporating SAM2 [45] masks, foreground segmentation, and open-world boxes and masks. The foreground segmentation, derived from the RMBG [74], supports diverse tasks such as inpainting and foreground extraction. Open-world bounding boxes are generated through the grounding caption capability of Qwen2-VL [56], which are processed using SAM2 [45] to produce corresponding masks.

Style transfer. We transfer the style of images according to reference in both semantic-variant and semantic-invariant settings. Specifically, the semantic-invariant transfer adopts InstantStyle [55] to preserve the semantic content, while the semantic-variant transfer relies on FLUX.1-Redux-dev [30], using the style embeddings and depth as conditions. For each image, we randomly generate five stylized versions. Mixing the two tasks pushes the model to better follow the in-context examples to avoid ambiguity.

Image editing. We design two types of editing tasks, including background-variant and background-invariant editing. The background-invariant editing begins with localizing the subjects. Then, we leverage a large vision-language model, Qwen2-VL [56], to modify the image caption with a new object that replaces the original subject. The image, with the subject masked, is subsequently processed by the FLUX.1-Fill-dev [30] inpainting model to integrate the alternative object into the masked region. The above operation is repeated five times to enrich the dataset. For background-variant editing, the difference lies on the last step, which utilizes FLUX.1-Redux-dev [30] with depth as the condition and the modified caption as the text prompt.

4. Method

This paper identifies the core challenges in building a universal image generation model, including the need for a clearly defined and generalizable task formulation, the sparsity of visual tasks, and the lack of a unified framework for multi-task learning. In the previous section, we addressed the issue of task sparsity by constructing the compact Graph200K dataset. Section 4.1 introduces visual in-context learning as the ideal paradigm for universal task formulation. Afterward, Section 4.2 considers the image infilling model as a unified multi-task framework, achieving strong generalization capabilities with minimal cost.

4.1. Visual In-context Learning

Language instructions are usually used to specify the generation definition to handle multiple visual generation tasks with a single generative model. However, due to the gap between vision and language, the text comprehension ability of image generation models remains limited. This issue leads to task confusion [35] in existing universal generative models and weak generalization to unseen tasks. Inspired by the success of few-shot learning on large language models [5], we recognize that visual context may serve as a more friendly task instruction for visual generative models, given their superior visual understanding capabilities.

Therefore, in this paper, we re-propose visual in-context learning to build a universal and generalizable image generation system. Formally, we define the image input-output of arbitrary conditional generation task as a query consisting of $L-1$ condition images and a blank target $\varnothing$ to be completed by the model, *i.e.*, $X = \mathrm{concat}(\{x_1, \ldots, x_{L-1}, \varnothing\})$. During training, we randomly provide 0 to C demonstration contexts ($X = \mathrm{concat}(\{x_1, \ldots, x_{L-1}, x_L\})$) for each query. This strategy ensures the generalization ability of models across different numbers of contexts while also preserving its capability to infer solely based on text instructions. As a result, the model can handle various applications, *i.e.*, users can efficiently process in-domain simple tasks using only text instructions or prompt the model with multiple visual contexts to execute complex and even unseen tasks. In our experiments, we show that providing task demonstrations not only helps alleviate task confusion and boost model performance across in-domain tasks [35], but also enhances the generalization ability on unseen tasks.

4.2. Unified Multi-task Framework

Unlike previous visual in-context learning models that primarily focus on scenarios with a single image condition and a single context [39, 52], in this work, we aim to construct a unified framework capable of handling varying numbers of conditions and contexts, allowing for flexible adaptation to diverse tasks. Assuming all images processed by the model share the same size $W \times H$, the input-output of our model naturally forms an interface of size $(L \times W) \times ((C+1) \times H)$ under our visual in-context learning formulation. Specifically, suppose we concatenate all inputs and outputs from a single interaction into a complete grid image (as shown in Fig. 1). In that case, the model we seek infills the last grid based on the surrounding context, akin to solving *visual cloze* puzzles. Therefore, we build our unified framework, **VisualCloze**, based on the general image infilling architecture capable of handling multiple resolutions.

Consistent with common diffusion-based infilling model designs, our model can be formulated as:

$$\hat{X} = f(X \mid T, M), \tag{1}$$

where X is the concatenated image, with the last grid left blank, T is the language instruction, M is the mask condition, and $\hat{X}$ represents the infilled result. The target image can be simply cropped from $\hat{X}$. The mask M is a binary matrix with the size of $(H \times (C+1), W \times L)$:

$$M(i,j) = \begin{cases} 1 & \text{if } i \in [H \times (C-1), H \times C) \\ & \text{and } j \in [W \times (L-1), W \times L)\,, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $M(i,j) = 1$ indicates that the pixel will be masked and generated by the infilling model. Specifically, we mask the region in the last row and column.

One key benefit of this design is that it shares a highly consistent optimization objective with general image infilling models without architectural modifications or explicit input conditions. This allows us to directly fine-tune advanced image infilling models using the newly constructed dataset while maximizing the utilization of the prior knowledge of foundation models. In contrast, existing task-specific models often require additional learnable modules [34, 60] or adapting to extra condition inputs [53], which may compromise the native capabilities of the model.

Note that the design of language instruction is also necessary for VisualClozebecause it is responsible for defining the grid image layout and specifying the task content when example contexts are unavailable. In our unified framework, the instruction consists of three parts: (1) layout instruction, which describes the $(C+1) \times W$ layout of the grid image, (2) task instruction, which specifies the task type, and (3) content instruction, which describes the content of the target image. The details about the instructions are available in the supplementary material. By restructuring the three components X, T, and M in Eqn. (1), we achieve a unified multi-task framework for image generation with the general image infilling paradigm and support in-context learning.

4.3. Implementation Details

We use FLUX.1-Fill-dev [30] as our foundation model, considering its outstanding performance among open-source

Condition	Method	Context	Controllability		Quality				Text Consistency
			F1 ↑	RMSE ↓	FID [22] ↓	SSIM ↑	MAN-IQA [65] ↑	MUSIQ [27] ↑	CLIP-Score [44] ↑
Canny	ControlNet [70]		0.13	-	46.06	0.34	0.31	45.45	34.10
	OminiControl [53]		0.47	-	29.58	0.61	0.44	61.40	34.40
	OneDiffusion [32]		0.39	-	32.76	0.55	0.46	59.99	34.99
	OmniGen [61]		**0.43**	-	51.58	0.47	0.47	62.66	33.66
	Ours_{dev}	0	0.39	-	**30.36**	**0.61**	0.48	61.13	**35.03**
	$\text{Ours}_{\text{fill}}$	0	0.35	-	30.60	0.55	**0.49**	**64.39**	34.98
	$\text{Ours}_{\text{fill}}$	1	0.36	-	31.34	0.55	**0.49**	64.12	34.96
	$\text{Ours}_{\text{fill}}$	2	0.36	-	31.15	0.56	**0.49**	64.08	34.85
Depth	ControlNet [70]		-	23.70	36.83	0.41	0.44	60.17	34.49
	OminiControl [53]		-	21.44	36.23	0.52	0.44	60.18	34.08
	OneDiffusion [32]		-	10.35	39.03	0.49	**0.49**	60.49	34.71
	OmniGen [61]		-	15.07	86.08	0.26	**0.49**	**64.90**	29.72
	Ours_{dev}	0	-	25.06	42.14	0.53	0.46	58.95	34.80
	$\text{Ours}_{\text{fill}}$	0	-	10.31	**33.88**	**0.54**	0.48	64.85	**35.10**
	$\text{Ours}_{\text{fill}}$	1	-	9.91	34.44	**0.54**	**0.49**	64.32	34.95
	$\text{Ours}_{\text{fill}}$	2	-	**9.68**	34.88	**0.54**	0.48	64.29	34.89
Deblur	ControlNet [70]		-	37.82	53.28	0.49	0.45	61.92	33.80
	OminiControl [53]		-	19.70	26.17	0.85	0.45	60.70	34.53
	OneDiffusion [32]		-	-	-	-	-	-	-
	OmniGen [61]		-	-	-	-	-	-	-
	Ours_{dev}	0	-	**25.03**	56.76	0.74	0.38	46.68	33.52
	$\text{Ours}_{\text{fill}}$	0	-	26.53	40.59	0.74	0.46	59.62	34.56
	$\text{Ours}_{\text{fill}}$	1	-	25.87	36.93	**0.76**	**0.48**	61.58	**34.82**
	$\text{Ours}_{\text{fill}}$	2	-	25.57	**36.28**	**0.76**	**0.48**	**61.77**	**34.82**

Table 1. Quantitative comparison on conditioning generation and image restoration. The methods that train a specialist for each task are marked as gray color. Except for these methods, the best method is bolded, and the second-best method is underlined.

Method	Context	DINOv2	CLIP-I	CLIP-T
OminiControl [53]		73.17	87.70	33.53
OneDiffusion [32]		73.88	86.91	34.85
OmniGen [61]		67.73	83.43	34.53
Ours_{dev}	0	78.05	87.68	35.06
$\text{Ours}_{\text{fill}}$	0	**80.41**	**89.63**	**35.16**
$\text{Ours}_{\text{fill}}$	1	79.33	89.22	35.02
$\text{Ours}_{\text{fill}}$	2	80.32	89.36	35.01

Table 2. Quantitative comparison for subject-driven image generation. We report clip scores on text alignment and style consistency. Specialist models for each task are shaded in gray. Among the remaining methods, the best one is emphasized in bold, while the second-best is underlined.

	text↑	image↑
InstantStyle [55]	0.27	0.60
OmniGen [61]	0.27	0.52
Ours_{dev}	**0.30**	0.53
$\text{Ours}_{\text{fill}}$	0.29	**0.55**

Table 3. Quantitative comparison for style transfer. We report CLIP scores on text alignment and style consistency. The specialist are indicated in gray. Among the others, the top-performing one is highlighted in bold, and the second-best is underlined.

models. We fine-tune the model using LoRA [24] with rank of 256. The model is tuned for 20,000 iterations with an accumulated batch size of 64 on 8 × A100 GPUs. We employ the AdamW optimizer with a learning rate of $1e^{-4}$. Following [30], we incorporate the lognorm noise strategy with dynamic time shifting. The number of contexts is set between 0 and 2, while the data stream length L varies between 2 and 4 in the Graph200K dataset. To balance computational efficiency, each grid size is fixed at 384 × 384. Although the generated resolution is relatively low, it is sufficient to validate the effectiveness of our proposed method. Moreover, in practical applications, high-resolution outputs can be obtained through simple post-upscaling techniques [41].

5. Experiments

5.1. Main Results

We compare our method with universal generative models, including OmniGen [61] and OneDiffusion [32], as well as specialized models, such as ControlNet [70] and OminiControl [53]. Additionally, we fine-tune FLUX.1-dev [30] using the same settings as FLUX.1-Fill-dev for comparison, and refer to the tuned models as Ours_{dev} and $\text{Ours}_{\text{fill}}$.

For conditional generation and image restoration, we evaluate the models based on three criteria, *i.e.*, controllabil-

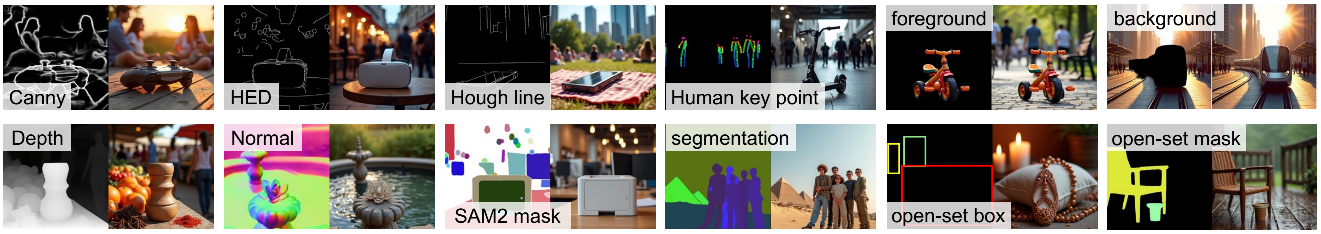


Figure 4. Illustration of conditional image generation.

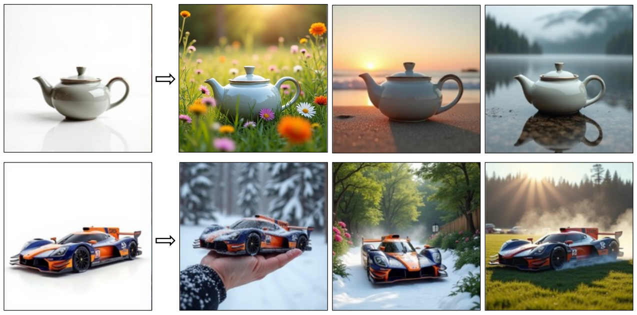


Figure 5. Subject-driven image generation. Based on a reference, we generate multiple images under different environments.

ity, visual quality, and text consistency, following the evaluation approach of OminiControl [53]. As shown in Tab. 1, our framework demonstrates comparable controllability to existing universal methods while achieving superior visual quality and text consistency. When compared to specialized methods, our model also performs on par with the best results and even outperforms them on the depth-to-image.

In the style transfer task, we measure text consistency and style alignment using the CLIP [44] model. As reported in Tab. 3, our method outperforms OmniGen [61] by 2% and 3% in text alignment and style consistency, respectively. Even when compared with InstantStyle-Plus [71], a specialized model, we achieve a 2% improvement in text consistency, with only a slight decrease in style alignment.

Furthermore, we evaluate the models on subject-driven image generation and report semantic alignment using the DINOv2 [42], CLIP-I [44], and CLIP-T [44] scores. Across all these metrics, our method consistently delivers improvements, as shown in Tab. 2. For example, compared to the specialized model, OminiControl [53], we achieve improvements of 7.15%, 1.66%, and 1.48% in these three scores.

Advantages of the infilling model. Our method (Ours$_{\text{fill}}$) is built on FLUX.1-Fill-dev [30], which shares the same objective as our unified image generation framework. To verify its effectiveness, we also fine-tune Fill.1-dev [30] (Ours$_{\text{dev}}$) using identical settings. Unlike Ours$_{\text{fill}}$, which requires no modifications, Ours$_{\text{dev}}$ necessitates model adaptations for universal image generation, as shown in the supplementary material. Despite its simplicity, Ours$_{\text{fill}}$ achieves superior performance across multiple tasks.

As shown in Tab. 1, Ours$_{\text{dev}}$ achieves a higher F1 score than Ours$_{\text{fill}}$ in canny-to-image generation. However, in other tasks, Ours$_{\text{fill}}$ demonstrates a significant advantage. For instance, in depth-to-image generation, Ours$_{\text{fill}}$ reduces RMSE from 25.06 to 10.31, indicating improved structural alignment. In the deblurring task, Ours$_{\text{fill}}$ achieves superior quality by lowering RMSE while maintaining a higher visual quality. In subject-driven image generation, Tab. 2 shows that Ours$_{\text{fill}}$ consistently outperforms Ours$_{\text{dev}}$. Additionally, in semantic-invariant style transfer, Ours$_{\text{fill}}$ delivers comparable performance to Ours$_{\text{dev}}$, as shown in Tab. 3.

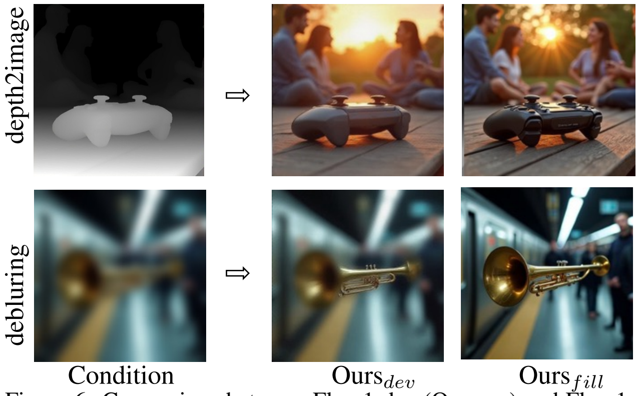


Figure 6. Comparison between Flux.1-dev (Ours$_{dev}$) and Flux.1-Fill-dev (Ours$_{fill}$).

Furthermore, Fig. 6 presents a visual comparison, where Ours$_{\text{fill}}$ demonstrates clear advantages over Ours$_{\text{dev}}$. Notably, in the depth-to-image generation, images produced by Ours$_{\text{dev}}$ frequently exhibit diagonal streak artifacts, which significantly degrade visual fidelity. Considering the advantages in performance, visual quality, and architectural efficiency, Ours$_{\text{fill}}$ stands out as the superior model.

5.2. Effectiveness of In-Context Learning

In this section, we analyze the impact of in-context learning on both seen and unseen tasks.

Quantitative comparison. Tab. 1 demonstrates the impact of in-context learning on different image generation tasks. Under the canny condition, our method without in-context examples achieves an FID of 30.60, which improves to 31.15 with two in-context examples. When conditioned on depth, the RMSE decreases from 10.31 to 9.68 as the number of in-context examples increases, indicating enhanced structural consistency. Similarly, in the deblurring task,

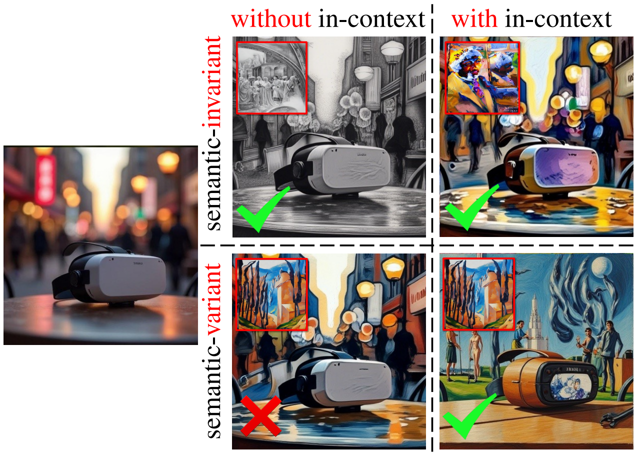


Figure 7. Effect of in-context learning on style transfer, best to zoom in to view. The images in the red box mean style references. Without in-context learning, semantic-variant style transfer (bottom) does not change the semantics as expected.

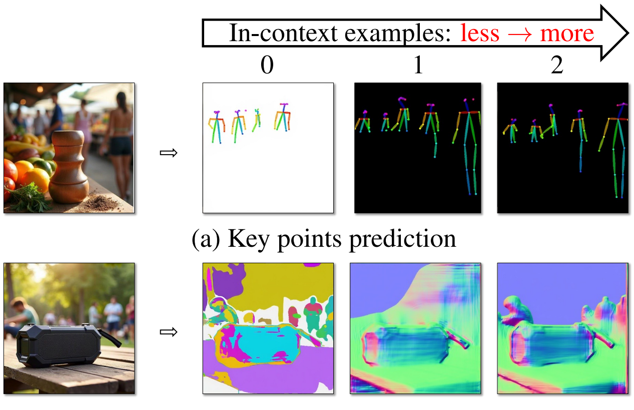
RMSE decreases from 26.53 to 25.57, reflecting improved fidelity to the original content. These results highlight in-context learning as an effective guidance mechanism, enabling the model to better align with the task intent.

In-domain task ambiguity. We observe that the model occasionally experiences task confusion, failing to accurately interpret the intended objective. In-context learning effectively alleviates this issue by providing task-specific demonstrates. As shown in Fig. 7, without in-context examples, the model struggles to distinguish between semantic-invariant and semantic-variant style transfers, leading to incorrect results where the background and semantics remain unchanged in semantic-variant cases. However, incorporating just a single in-context example enables the model to correctly perform semantic-variant style transfer. Similarly, as shown in Fig. 8, increasing the number of in-context examples enhances the performance. Additional visualizations are provided in the supplementary materials.

Generalization to unseen tasks. Beyond mitigating task confusion, in-context learning also improves the ability of models to generalize to tasks unseen during training. As shown in Fig. 9, we evaluate the performance in generating frontal faces from side-view images, a task not encountered during training. Without in-context examples, the generated faces fail to align with the expected frontal view. However, as the number of in-context examples increases, the model progressively refines its outputs, producing more accurate frontal faces. These results demonstrate that in-context learning serves as an effective guidance mechanism, enabling adaptation to novel tasks without retraining.

6. Conclusion

In this work, we propose VisualCloze, a universal image generation framework that addresses key challenges in existing methods, including generalizable instruction design, appropriate task distributions, and unified architectural design. Rather than relying solely on language-based instructions to convey task intent, we re-propose visual in-context learning, enabling the model to learn tasks from demonstrations. This approach improves generalization to unseen tasks and reduce task ambiguity. To overcome the sparsity of visual task distributions, which limits the learning of transferable knowledge, we construct Graph200K, a graph-structured dataset that establishes interrelated tasks. In this compact task space, the model is promoted to learn transferable representations and improve adaptability. Meanwhile, we identify the consistent optimization objective between image infilling and universal generation, allowing us to seamlessly adapt general-purpose infilling models for universal generation without architectural modifications. Experimental results show that our approach supports a diverse set of in-domain tasks using in-context learning while demonstrating strong generalization to unseen tasks.



(a) Key points prediction

(b) Normal map estimation

Figure 8. Effect of in-context learning on dense prediction.

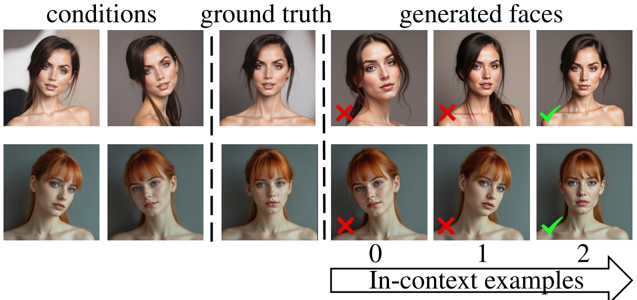


Figure 9. Effect of in-context learning on the unseen task: generating frontal face conditioning on side faces. More in-context examples produce a more accurate frontal face.

Acknowledgments. This work is supported by the Shenzhen Science and Technology Program (JCYJ20240813114237048), and the National Natural Science Foundation of China (Grant No. 62206272).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 2, 3

[2] Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR*, 2023. 3

[3] Ivana Balazevic, David Steiner, Nikhil Parthasarathy, Relja Arandjelovic, and Olivier J Henaff. Towards in-context scene understanding. In *NeurIPS*, 2023. 2, 3

[4] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A Efros. Visual prompting via image inpainting. In *NeurIPS*, 2022. 2, 3

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020. 2, 3, 5

[6] John Canny. A computational approach to edge detection. *IEEE TPAMI*, 1986. 4

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE TPAMI*, 2019. 4

[8] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, Hui Ding, Zhe Lin, and Hengshuang. Unireal: Universal image generation and editing via learning real-world dynamics. *arXiv preprint arXiv:2412.07774*, 2024. 3

[9] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018. 2

[10] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, 2021. 1

[11] Zheng Chong, Xiao Dong, Haoxiang Li, shiyue Zhang, Wenqing Zhang, Hanqing Zhao, xujie zhang, Dongmei Jiang, and Xiaodan Liang. CatVTON: Concatenation is all you need for virtual try-on with diffusion models. In *ICLR*, 2025. 1

[12] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021. 3

[13] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2024. 3

[14] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024. 1, 3

[15] Christopher Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. In *NeurIPS*, 2021. 2

[16] Peng Gao, Le Zhuo, Dongyang Liu, Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, Chen Lin, Rongjie Huang, Shijie Geng, 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] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Mdtv2: Masked diffusion transformer is a strong image synthesizer. *arXiv preprint arXiv:2303.14389*, 2024. 3

[18] Golnaz Ghiasi, Barret Zoph, Ekin D. Cubuk, Quoc V. Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In *ICCV*, 2021. 2

[19] Geonmo Gu, Byungsoo Ko, SeoungHyun Go, Sung-Hyun Lee, Jingeun Lee, and Minchul Shin. Towards light-weight and real-time line segment detection. In *AAAI*, 2022. 4

[20] Aaron Hertzmann. *Algorithms for rendering in artistic styles*. PhD thesis, New York University, Graduate School of Arts and Science, 2001. 3

[21] Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. Image analogies. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, 2001. 3

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 6

[25] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Huanzhang Dou, Yupeng Shi, Yutong Feng, Chen Liang, Yu Liu, and Jingren Zhou. Group diffusion transformers are unsupervised multitask learners. *arXiv preprint arxiv:2410.15027*, 2024. 3

[26] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arxiv:2410.23775*, 2024. 1, 2

[27] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 6

[28] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 2

[29] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, 2017. 2

[30] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024. 1, 2, 3, 4, 5, 6, 7

[31] Bolin Lai, Felix Juefei-Xu, Miao Liu, Xiaoliang Dai, Nikhil Mehta, Chenguang Zhu, Zeyi Huang, James M Rehg, Sangmin Lee, Ning Zhang, et al. Unleashing in-context learning of autoregressive models for few-shot image manipulation. *arXiv preprint arXiv:2412.01027*, 2024. 3

[32] Duong H. Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and Jiasen Lu. One diffusion to generate them all, 2024. 1, 2, 3, 6

[33] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE TPAMI*, 2023. 4

[34] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, 2024. 1, 3, 5

[35] Weifeng Lin, Xinyu Wei, Renrui Zhang, Le Zhuo, Shitian Zhao, Siyuan Huang, Junlin Xie, Yu Qiao, Peng Gao, and Hongsheng Li. Pixwizard: Versatile image-to-image visual assistant with open-language instructions. *arXiv preprint arXiv:2409.15278*, 2024. 1, 5

[36] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. 3

[37] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024. 3

[38] 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

[39] Yihao Liu, Xiangyu Chen, Xianzheng Ma, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Unifying image processing as visual prompting question answering. *arXiv preprint arXiv:2310.10513*, 2023. 2, 3, 5

[40] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487*, 2025. 1, 3

[41] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 6

[42] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7

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

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6, 7

[45] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4

[46] 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 (CVPR)*, pages 10684–10695, 2022. 3

[47] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017. 2

[48] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 1

[49] Dianmo Sheng, Dongdong Chen, Zhentao Tan, Qiankun Liu, Qi Chu, Jianmin Bao, Tao Gong, Bin Liu, Shengwei Xu, and Nenghai Yu. Towards more unified in-context visual understanding. In *CVPR*, 2024. 3

[50] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 3

[51] Yanpeng Sun, Qiang Chen, Jian Wang, Jingdong Wang, and Zechao Li. Exploring effective factors for improving visual in-context learning. *arXiv preprint arXiv:2304.04748*, 2023. 3

[52] Yasheng SUN, Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, and Hideki Koike. Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. In *NeurIPS*, 2023. 2, 3, 5

[53] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 3, 2024. 3, 4, 5, 6, 7

[54] Alex Jinpeng Wang, Linjie Li, Yiqi Lin, Min Li, Lijuan Wang, and Mike Zheng Shou. Leveraging visual tokens for extended text contexts in multi-modal learning. *NeurIPS*, 2024. 3

[55] Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788*, 2024. 1, 3, 4, 6

[56] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's

perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 4

[57] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, 2023. 2, 3

[58] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *ICCV*, 2023. 2, 3

[59] Zhendong Wang, Yifan Jiang, Yadong Lu, yelong shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. In *NeurIPS*, 2023. 3

[60] Cong Wei, Zheyang Xiong, Weiming Ren, Xinrun Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. *arXiv preprint arXiv:2411.07199*, 2024. 1, 2, 5

[61] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. 1, 2, 3, 6, 7

[62] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *CVPR*, 2015. 4

[63] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE TPAMI*, 2023. 4

[64] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 4

[65] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1191–1200, 2022. 6

[66] 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

[67] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. *arXiv preprint arXiv:2401.13627*, 2024. 2

[68] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. 3

[69] Hayoung Yun and Hanjoo Cho. Achievement-based training progress balancing for multi-task learning. In *ICCV*, 2023. 2

[70] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 4, 6

[71] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *CVPR*, 2023. 1, 2, 7

[72] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? In *NeurIPS*, 2023. 2, 3

[73] Canyu Zhao, Mingyu Liu, Huanyi Zheng, Muzhi Zhu, Zhiyue Zhao, Hao Chen, Tong He, and Chunhua Shen. Diception: A generalist diffusion model for visual perceptual tasks. *arXiv preprint arXiv:2502.17157*, 2025. 3

[74] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 2024. 4

[75] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 2

[76] Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. *arXiv preprint arXiv:2402.11574*, 2024. 3

[77] Muzhi Zhu, Yang Liu, Zekai Luo, Chenchen Jing, Hao Chen, Guangkai Xu, Xinlong Wang, and Chunhua Shen. Unleashing the potential of the diffusion model in few-shot semantic segmentation. In *NeurIPS*, 2024. 2, 3

[78] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, Xu Luo, Zehan Wang, Kaipeng Zhang, Lirui Zhao, Si Liu, Xiangyu Yue, Wanli Ouyang, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-next : Making lumina-t2x stronger and faster with next-dit. In *NeurIPS*, 2024. 1, 2, 3