Instruct-Imagen: Image Generation with Multi-modal Instruction

Hexiang Hu[♠⋆] Kelvin C.K. Chan[♢⋆] Yu-Chuan Su[♢⋆] Wenhu Chen[♠⋆]
Yandong Li[♢] Kihyuk Sohn[♢] Yang Zhao[♢] Xue Ben[♢] Boqing Gong[♢]
William Cohen[♠] Ming-Wei Chang[♠] Xuhui Jia[♢]
[♠]Google DeepMind [♢]Google Research
{hexiang,kelvinckchan,ycsu,wenhuchen}@google.com

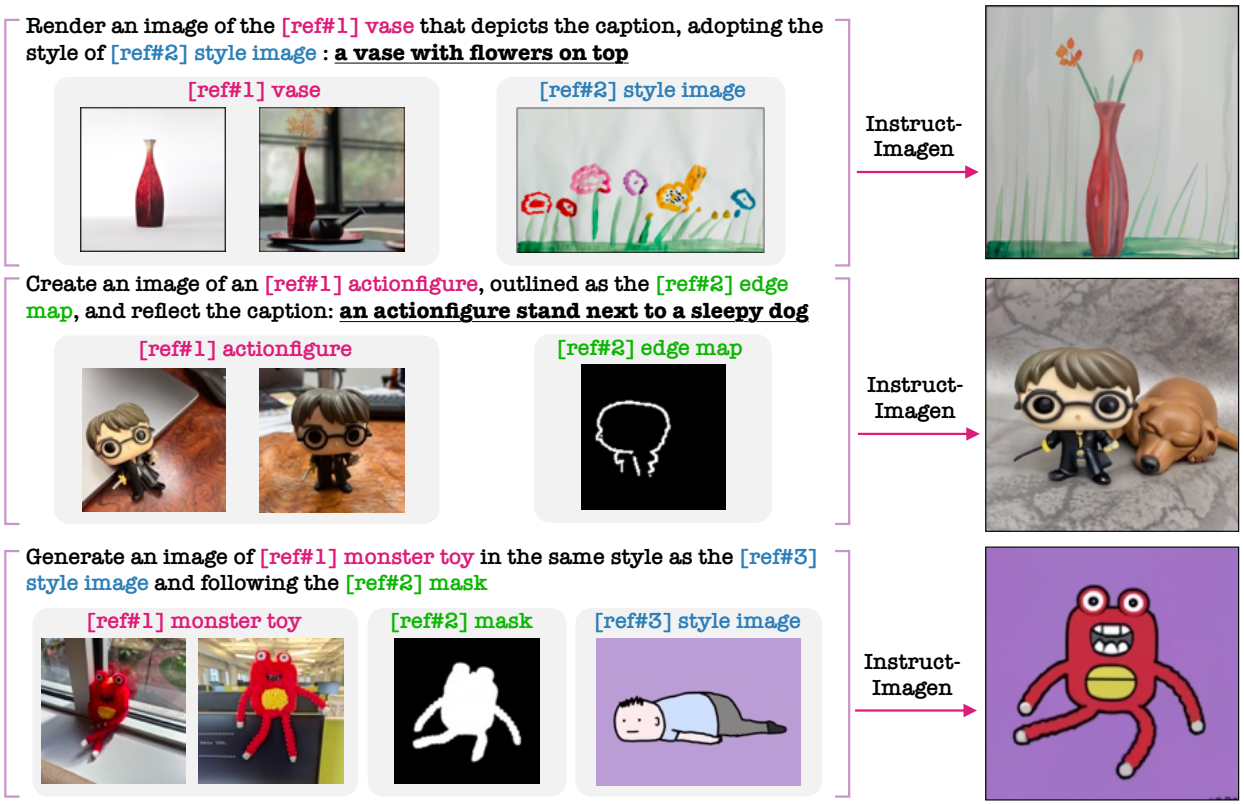


Figure 1. **Zero-shot generalization of Instruct-Imagen**. Our model understands the multi-modal instruction (left) to generate image (right) that reflects the complex and unseen image transformation.

Abstract

This paper presents Instruct-Imagen, a model that tackles heterogeneous image generation tasks and generalizes across unseen tasks. We introduce **multi-modal instruction** *for image generation, a task representation articulating a range of generation intents with precision. It uses natural language to amalgamate disparate modalities (*e.g.*, text, edge, style, subject,* etc.*), such that abundant generation intents can be standardized in a uniform format.*

*We then build Instruct-Imagen by fine-tuning a pre-trained text-to-image diffusion model with two stages. First, we adapt the model using the retrieval-augmented training, to enhance model's capabilities to ground its generation on external multi-modal context. Subsequently, we fine-tune the adapted model on diverse image generation tasks that requires vision-language understanding (*e.g.*, subject-driven generation,* etc.*), each paired with a multi-modal instruction encapsulating the task's essence. Human evaluation on various image generation datasets reveals that Instruct-Imagen matches or surpasses prior task-specific models in-domain and demonstrates promising generalization to unseen and more complex tasks. Our evaluation suite will be made publicly available.*

⋆ These authors contributed equally to this work.

1. Introduction

The advent of generative artificial intelligence (GenAI) has ushered in an era of significant advancements in image generation, primarily through text-to-image models. Existing models such as Stable Diffusion [35], DreamBooth [37], StyleDrop [42], ControlNet [50] mainly focus on accepting specific instruction modality like text prompt, subject, style, edge, *etc*. Their ability to comprehend more complex instructions involving multiple modalities (*e.g*., subject + mask + style) is yet to show, not to mention its ability to generalize to unseen instructions [20].

Unlike the language generation [2, 11, 27, 27, 45], image generation inherently involves multimodality. In the realm of human artistry, the painting process often integrates various modalities to achieve the desired outcome. A painter might start with a rough sketch to outline the composition, then apply a specific style, like impressionism, for details on texture and color. They may also use photographs or live models as subject references, blending these elements to create an expressive piece of art. Communicating the multi-modal complexities behind such an "image generation" procedure is challenging, even among humans.

Can we effectively communicate the multi-modal complexities to models? To address this challenge, we introduce *multi-modal instruction* in image generation. This approach interleaves and adheres information from different modalities, expressing the conditions for image generation (refer to Figure 1 left for examples). Specifically, multi-modal instruction enhances language instructions, *i.e*., "render an instance of `subject images` adopting the style of `style image`, such that...", by integrating information from other modalities (*e.g*., subject and style) to describe the objective of generating a customized image of the given subject in the provided visual style. As such, prior image generation tasks with multi-modal conditions can be efficiently communicated in a human intuitive interface (see § 2).

We then build our model, *i.e*., `Instruct-Imagen`, employing a two-stage training approach, to first enhance model's ability to process multi-modal instructions, and then faithfully follow the multi-modal user intents. This involved initially adapting a pre-trained text-to-image model to handle additional multi-modal inputs, followed by fine-tuning it to accurately respond to multi-modal instructions. Particularly, we begin by continuing the text-to-image generation training of a pre-trained diffusion model, supplemented by similar `(image, text)` contexts retrieved from a web-scale corpus [6]. In the second stage, we fine-tune the model on diverse image generation tasks, each paired with multi-modal instructions that encapsulate the essence of the task. Consequently, `Instruct-Imagen` excels in merging diverse modal inputs like sketches and visual styles with textual directives, producing contextually accurate and visually compelling images.

As illustrated in Figure 1, `Instruct-Imagen` demonstrates strong capability of understanding the sophisticated multi-modal instruction to generate the images faithful to the human intention, even when the instruction combination has never been observed before. Human studies establishes that `Instruct-Imagen` not only matches but, in several instances, surpasses prior task-specific models within their domains. More significantly, it exhibits a promising generalization capability when applied to unseen and more complex image generation tasks.

We summarize our contributions as follows:

- We introduce multi-modal instruction, a task representation universally represents instruction from multiple modalities, *e.g*., text, edge, mask, style, subject, *etc*.
- We propose to perform retrieval-augmented training and multi-modal instruction-tuning to adapt the pre-trained text-to-image models to follow multi-modal instructions.
- We build `Instruct-Imagen`, a unified model that tackles heterogeneous image generation tasks, surpassing the several state-of-the-arts in their domains.
- More substantially, `Instruct-Imagen` generalizes to unseen and complex tasks, *without any ad hoc design*.

2. Multi-modal Instructions for Generation

In this section, we start with discussing the preliminary on diffusion models with input conditions. Then we introduce the format of multi-modal instruction, and discuss how prior image generation tasks can be unified in this framework.

Diffusion Models with Input Conditions. Diffusion models [35, 38, 41] are latent variable models, parameterized by Θ, in the form of $p_\Theta(\boldsymbol{x}_0) := \int p_\Theta(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T}$, where $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T$ are "noised" latent versions of the input image $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$. Note that the dimension of both latent and the image are the same throughout the entire process, with $\boldsymbol{x}_{0:T} \in \mathbb{R}^d$ and d indicating the data dimension. The process that computes the posterior distribution $q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$ is called the diffusion process, and is implemented as a predefined Markov chain that gradually adds Gaussian noise to the data according to a schedule β_t:

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}); \tag{1}$$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}) \tag{2}$$

Diffusion models are trained to learn the image distribution by reversing the diffusion Markov chain. Theoretically, this reduces to learning to denoise $\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ into $\boldsymbol{x}_0$, with a time re-weighted square error loss [15]:

$$\mathbb{E}_{(\boldsymbol{x}_0, \boldsymbol{c})\sim D}\{\mathbb{E}_{\boldsymbol{\epsilon}, t}[w_t \cdot ||\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}) - \boldsymbol{x}_0||_2^2]\} \tag{3}$$

where D is the training dataset containing (image, condition) = $(\boldsymbol{x}_0, \boldsymbol{c})$ pairs. In the text-to-image models, the con-

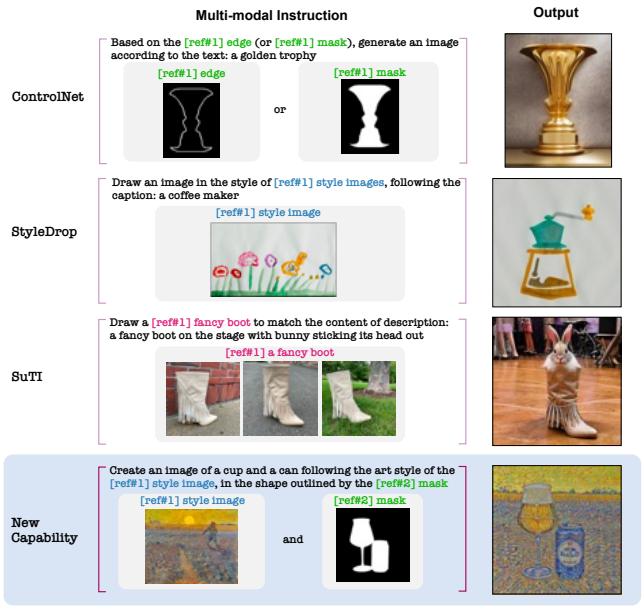


Figure 2. Illustration on how **multi-modal intruction** uniformly express existing image generation tasks and extends to new tasks. Examples in this figure are retrieved from [7, 42, 50]

dition $\boldsymbol{c}$ are often the embeddings of input text prompt, from pre-trained text embedding models (*e.g.*, T5 [32]).

Unified Multi-modal Instruction. While multi-modality information is necessary for extended image generation applications, and had been explored in prior works [7, 22, 37, 42, 50], *etc.*, there was not such a format in the literature that allows generalization. Instead, models often make ad-hoc design to integrate information from other modalities. For example, ControlNet [50] combines the input $\boldsymbol{x}_t$ with a transformed spatial control map feature to form the new input for reverse diffusion. Such modality and task specific design, while effective in-domain, is challenging to generalize to other tasks (*e.g.*, stylization). Therefore, we propose the **multi-modal instruction**, a new format where language are used to explicitly state the objective behind tasks, with references to multi-modal conditions.

There are two key components in the proposed instruction format: (1) the payload text instruction that provides detailed description of the task objective, with reference marker (*e.g.*, `[ref#?]`). (2) a *multi-modal context* with (marker + text, image) pairs. The model then employ a shared instruction understanding model to consume both the text instruction and the multi-modal context, regardless of the specific modality in the context. Figure 2 showcased three examples of how this format represents various prior generation tasks, showing its compatibility to prior image generation tasks. More importantly, the flexibility of language allows multi-modal instructions to extend to new tasks, without any modality & task specific design.

3. `Instruct-Imagen`

In this section, we first discuss how `Instruct-Imagen` encodes the input multi-modal instruction, and how the encoding is leveraged for generation (see § 3.1). Then we introduce the two staged training framework for `Instruct-Imagen` in § 3.2. In Figure 3, we present the high-level design of the `Instruct-Imagen`, alongside with an overview of its training procedure.

3.1. Imagen with Multi-modal Instruction

The foundation of `Instruct-Imagen` is the multi-modal instruction, which uniformly represents prior image generation tasks, while remains its capability to extend to novel and complex tasks. Based on it, we designed the model architecture that extends a pre-trained text-to-image diffusion models, *i.e.*, a cascaded diffusion model [16], to allow it fully conditioned on the input multi-modal instruction.

Cascaded Backbone Text-to-Image Model. We used a version of Imagen [38] pre-trained on internal data sources, which inherents the cascaded text-to-image diffusion model (see Figure 3 left), as the founding for adaptation to `Instruct-Imagen`. The full model has two sub-components: (1) a text-to-image that generates $128\times$ resolution images from text prompt only, and (2) a text-conditioned super-resolution model that scales the 128 resolution up to high fidelity $1024\times$ images. In the scope of this work, we only consider training and adapting the 128 resolution text-to-image network, for the sake of efficiency and clarity. Particularly, the backbone model is a convolutional UNet [36] with bottleneck, with a paired down-sampling encoder and up-sampling decoder. The text are then embedded with a pre-trained T5-XXL model [32]. The embeddings are then input to the down-sampling encoder as condition, and to the cross-attention on bottleneck representation as enhanced reference.

Encoding Multi-modal Instruction. We adapt the above mentioned cascaded text-to-image model via maximally reusing the pre-trained text-to-image model for encoding the multi-modal instruction, and only introduce one cross-attention layer that conditions the bottleneck representation of UNet with the embedded multi-modal context the (key, value) pairs. This grows the number of parameters of our model from $2.51B$ to $2.76B$ ($\sim$10%). This design is in principle similar to the nearest neighbor UNet presented in [6] (but with the nested encoding on the multi-modal context). Figure 3 (right) illustrates the dataflow of how a multi-modal instruction is encoded by the `Instruct-Imagen`. Here, the payload text instruction is encoded the same way as normal text input in backbone model. The multi-modal context, *i.e.*, both (marker + text, image) pairs, are first encoded using the down-sampling encoder, same as how backbone text-to-image model encodes the bottleneck represen-

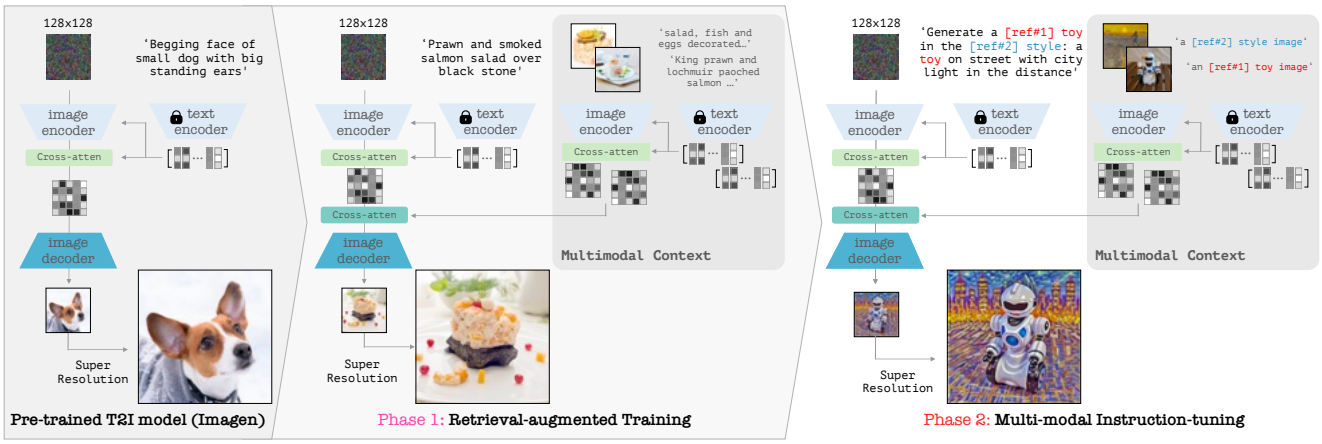


Figure 3. Overview of the two-staged training pipeline for the proposed `Instruct-Imagen` model.

tation, and then provided as (key, value) pairs for the new cross-attention layer to condition on. The up-sampling decoder then takes the outcome feature representation to perform the reverse diffusion.

3.2. Training `Instruct-Imagen` in Two Stages

Our training pipeline is two staged, with the first stage to continue the text-to-image generation, with augmentation of retrieved neighbor `(image, text)` pairs. Then in the second stage, we fine-tune the output model from first stage on a mixture of diverse image generation tasks, each paired with corresponding multi-modal instructions. In both training stages, the model are optimized end-to-end.

Retrieval-augmented Text-to-image Training. The most important research question for `Instruct-Imagen` is how to train the model to condition on multi-modal inputs for its generation, since these tasks deviate from the standard text-to-image pre-training. A straight-forward thinking is to mine naturally distributed multi-modal Internet data [1, 52] (such as Wikipedia articles with images) and train models to use the interleaved `(image, text)` data to generate the desired output image. However, this is inadequate to train models with superior alignment, because the input multi-modal content are often not relevant to the production of the output image. For example, in the Wikipedia article, *i.e.*, `the US president`, the headline text, summary text and info-box images (*i.e.*, `Biden`'s picture) are not informative to generate the image of `Franklin D. Roosevelt`. Thus, training model using such data often leads to ignorance of the multi-modal context.

To alleviate this issue, we employ the training data similar to re-imagen [6], such that the model can learn to look at the relevant but not duplicated neighboring multi-modal context when generating image according to the current text prompt. Particularly, the model would be presented with portraits of `Franklin D. Roosevelt` at other occurrences, when asked to generate his presence delivering the radio address in 1933. A model capable of processing multi-modal inputs can leverage other `Roosevelt` images to generate the scene, instead of memorizing his appearance.

Task	Input	Dataset	#Examples	Ratio
Txt2Img	`txt`	Internal Data	5M	0.15
		WikiArt	0.1M	0.05
Control2Img	`depth_img+txt`	Depth WebLI [8]	5.7M	0.06
	`mask_img+txt`	Mask WebLI [8]	5.7M	0.06
	`edge_img+txt`	Edge WebLI [8]	5.7M	0.06
		Sketch2Image [23]	15K	0.02
Subject Txt2img	`sub_imgs+txt`	SuTI dataset [7]	0.75M	0.30
		Celeb-A [25]	0.1M	0.05
		Celeb-HQ [19]	0.1M	0.05
Style Txt2img	`sty_img+txt`	Derived from WikiArt	0.1M	0.10
Style Transfer	`sty_img+ctn_img`	WikiArt + Internal Data	1M	0.10

Table 1. Details of the instruction-tuning datasets and mixing ratio.

To achieve this, we construct the retrieval-augmented training dataset via domain-specific clustering of Web `(image, text)` pairs. First, we processed the web scale image-text corpus (*i.e.*, WebLI [8, 9]) to remove low quality images (in image quality scores [43]), classified images from specific clusters (*e.g.*, art, products, animals, scenery, *etc.*) via image-text matching, and performed image clustering within each classified sub-cluster, using the embeddings from CLIP [31] model. For each mined image cluster, we took the top 10 nearest neighbor candidates, and performed near-duplication removal via removing images with high similarity and images with the same metadata (*e.g.*, URL). We then truncate the image cluster to have the size of 5 images (discarded clusters with less than 5 images). As an outcome, this process produced 8.08 M `(image, text)` clusters, with 5 pairs per cluster. During the training, one `(image, text)` pair is sampled as the input and target for the `Instruct-Imagen`, and three other `(image, text)` pairs are sampled as the multi-modal context. Additionally, we performed the condition dropout as [35, 38] but with two independent drop situations: (1) dropping both the input text and multi-modal context; and (2) dropping only the multi-modal context, each dropout situation occurs at 10% chance.

Multi-modal instruction-tuning for Image Generation.

We prepared 11 image generation datasets via either re-using existing dataset or synthesizing the input or target image, which formed 5 task categories, for multi-modal instruction-tuning. For each dataset, we prompted the GPT-4 [27] to generate 100 rephrased instruction templates with high variation, and validated the semantic correctness of them manually. We defer the qualitative examples of each dataset and its associated instruction to the appendix. The Table 1 presents the detailed information about task group, model input conditions, and data statistics for each prepared dataset, with details below:

- **Text-to-image Generation.** We processes two datasets for instructed text-to-image generation: an internal high-quality natural image dataset with manual caption; and an art specific dataset crawled from WikiArt (using the pipeline in [44]), with the caption generated by PaLI [8]. Both datasets are augmented with sampled instruction.
- **Control2Image Generation.** We followed [50] to prepare the control signals (*e.g.*, depth map, mask, and edge), based on a subset of the WebLI [8]. Specifically, we use MiDas [34] for depth estimation, HED [46] for edge extraction, and salient object [30] for mask. To improve robustness with different edge styles, we also employed edge-to-image data from a sketch dataset [23].
- **Subject-driven Generation.** We consider two data sources for subjects: general objects and human instances, for subject-driven generation. Particularly, we use the subject-driven dataset introduced in SuTI [7] for general object learning, and the celebrity face datasets [19, 25] to learn face rendering. For face rendering, we group the faces of the same person and caption them with PaLI [8], then we use one sampled example as the input/target, and the rest as multi-modal context. All datasets then join the instruction templates, with reference markers inserted to refer the multi-modal context.
- **Styled Generation.** Styled generation is a task that generalizes over the StyleDrop [42], with a style image and text as input, styled image following the text as output. To collect such data, we used images from WikiArt as the collection of style images to train StyleDrop models, and then use the manual captions from the internal text-to-image dataset to sample images as the target styled image. We employ a CLIP model to filter out examples that fails the alignment with either style image or the caption. Then multi-modal instructions are created via combining the instruction template with style image and the caption, such that the style image is correctly referred.
- **Style Transfer.** Similarly, we construct the style transfer dataset via combining style images from our WikiArt crawl and content images from the internal dataset (with the captions discarded). Particularly, we employ a simple style transfer model [13], which allows fast and large-scale generation, to blend the style image with the content image. These data are then augmented with instructions.

During the instruction-tuning stage, we fine-tune the output model of the retrieval-augmented training on the multi-task mixed dataset, with the mixture ratio specified in Table 1.

4. Related Work

Instruction-Tuning. Instruction tuning was first introduced in FLAN [45], which finetunes a large language model (LLM) on instructions to significantly improve its zero-shot learning performance on unseen tasks. Chung et al. extended the work at scale [11], showing extraordinary generalization to numerous NLP tasks. In general, the instruction data plays a pivotal role in the finetuned LLM [51]. This success experience in text instruction tuning was then introduced to the vision-language models [4, 9, 24], which enables generalization across tasks [10, 14, 17, 26].

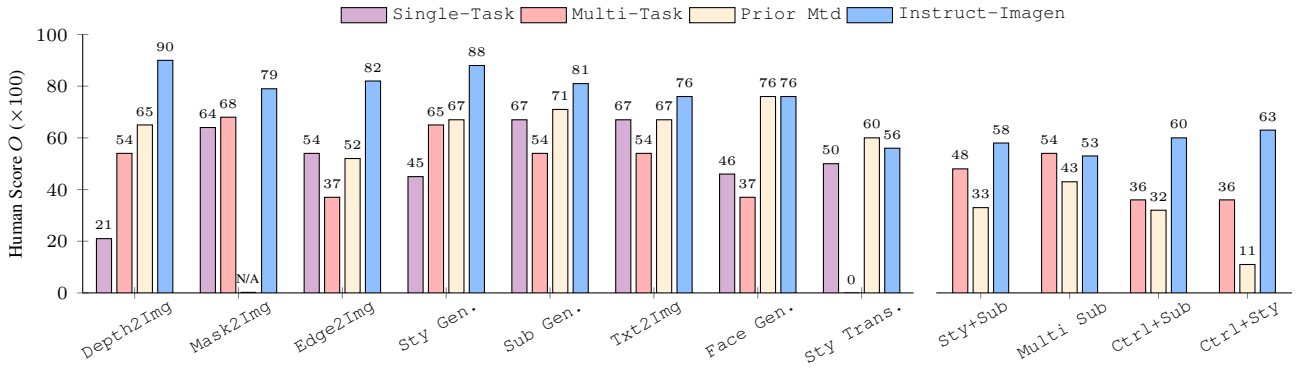
Controlled Image Synthesis. Recent advancements in text-to-image generative models [3, 5, 6, 33, 35, 38, 47, 48] have showcased impressive capabilities in various domains, including creativity, photorealism, diversity, and coherence. A critical aspect of these advancements is controllability, which has been enhanced by adapting these models to specific subjects [7, 37], styles [42], masks [50], *etc*. For example, DreamBooth [37] fine-tunes a text-to-image model on a limited set of images to better capture the nuances of a specific subject. Additionally, ControlNet [50] introduces the ability to condition on a variety of control signals, including depth maps and doodles, by fine-tuning an auxiliary encoder with the appropriate data pairs.

5. Experiments

In this section, we first introduce the experimental setup, the human evaluation protocol, and comparative baseline systems in § 5.1, then present the main results in § 5.2, and finally perform an in-depth analysis in § 5.3.

5.1. Experimental Setup

We evaluate our models with two setups, *i.e.*, *in-domain task evaluation* and *zero-shot task evaluation*, where the later setup is strictly more challenging than the former. Particularly, we re-use the recently introduced conditional image generation benchmark, *i.e.*, ImagenHub [20], for evaluating text-to-image generation. We also employ other datasets to cover in-domain evaluation: We adopt the DreamBench [7, 37] v1 & v2 as our subject-driven evaluation data; We use the style images from StyleDrop [42] for style evaluation; We use hold-out style images from WikiArt [44] and content images from CustomConcept101 [21] for style transfer. We use the evaluation data of WebLI [8] for control2image (*i.e.*, mask, edge, depth) evaluation. For face evaluation, we evaluate on the validation set of hold-out human in CelebA [25] and CelebA-HQ [19].



(a) In-domain Evaluation (b) Zero-shot Evaluation

Figure 4. **Human Study on prior methods, baselines, and `Instruct-Imagen`.** `Instruct-Imagen` can perform on par or better comparing to the baselines and prior methods, with best generalization capability to novel tasks. `Instruct-Imagen` does not require any fine-tuning for all tasks (particularly style/subject-related), and inferences at an average speed of 18.2 seconds per example (on TPUv4).

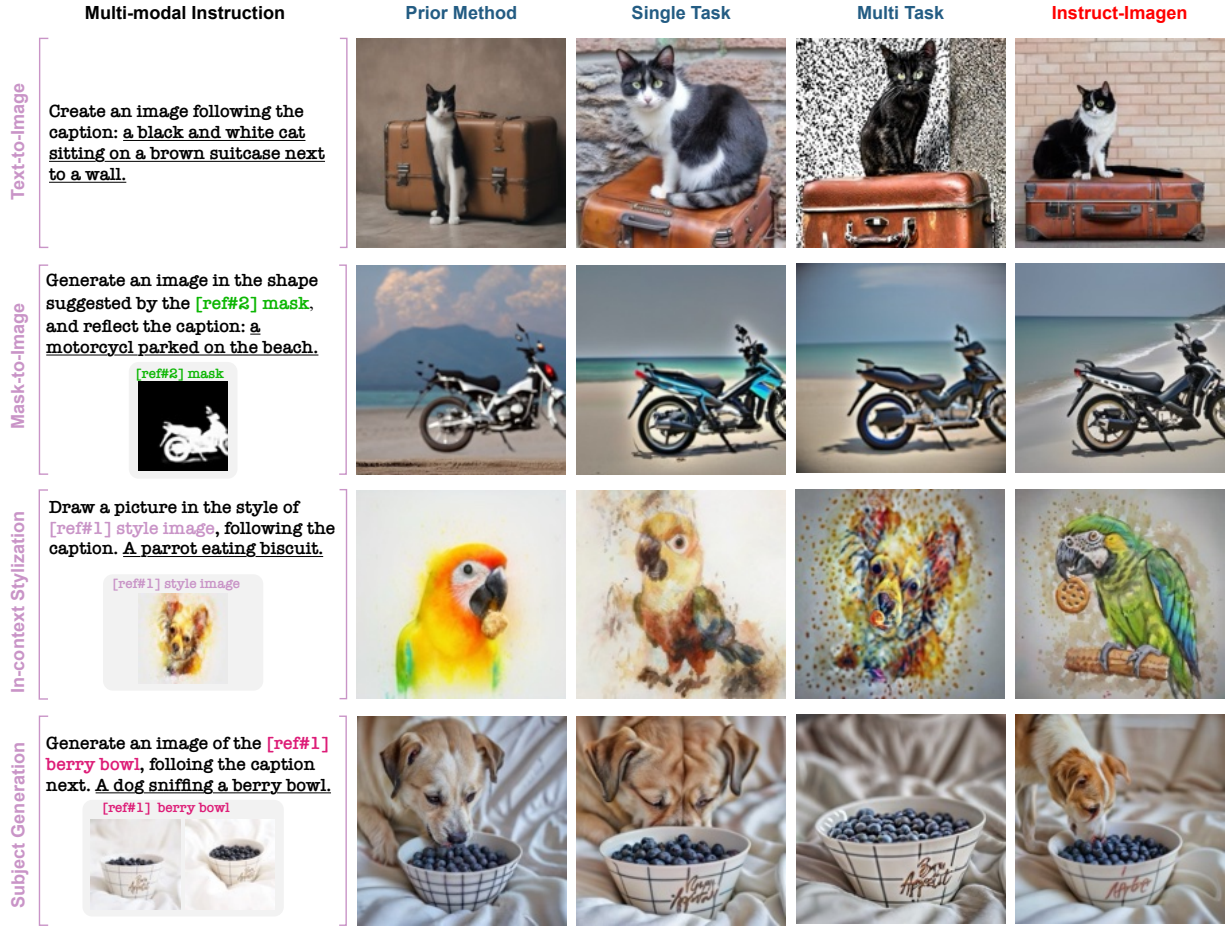


Figure 5. **Comparison on a subset of in-domain tasks.** Examples generated from prior methods, baselines, and `Instruct-Imagen`. We visualize the multi-modal instruction for human intuitive understanding (models are evaluated with in-distribution inputs).

For zero-shot tasks, we either adopt the existing evaluation (*i.e.*, multi-subject evaluation using a subset of 100 examples on CustomConcept101 [21]) or construct the evaluation ourselves (*e.g.*, subject + control, style + control, style + subject) by adopting images from corresponding datasets. For example, we use subject images from the CustomConcept101 [21] and style images from WikiArt [44] to perform subject + style evaluation. We refer the readers to the appendix for complete information about evaluation datasets. The complete evaluation suite would be made publicly available for future study and comparison.

Baseline Models. We compare `Instruct-Imagen` with three category of baseline models: (1) Prior State-of-the-art method (2) Single-task model (3) Multi-task model. Since

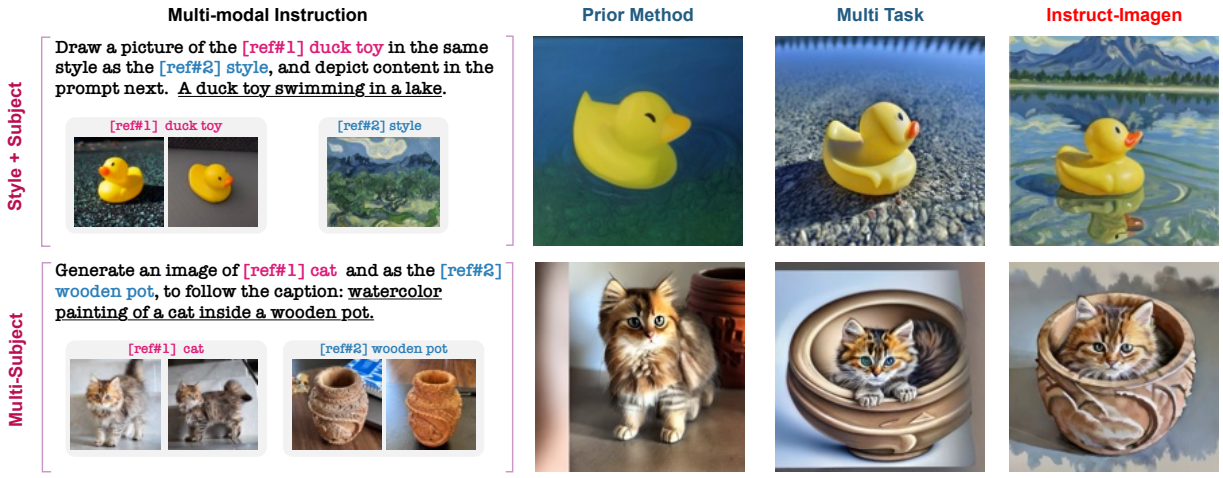


Figure 6. **Comparison on a subset of zero-shot tasks.** Examples generated from prior methods, the baseline, and `instruct-imagen`. We visualize the multi-modal instruction for human intuitive understanding (models are evaluated with in-distribution inputs).

no single prior model can handle all image generation tasks, we make comparison to different prior method on each task. Particularly, we compare to: SDXL [29] for text-to-image generation; ControlNet [50] for edge/depth-to-image generation; Ghiasi *et al*. [13] for style transfer; StyleDrop [42] for styled generation; SuTI [7] for subject-driven generation; and TamingEncoder [18] for face generation. Note that we marked prior method on `Mask2Img` task with `N/A` due to lack of public model. For zero-shot tasks, we compare to: StyleDrop+DreamBooth [37, 42] for styled subject generation; CustomDiffusion [21] for multi-subject generation; and KOSMOS-G [28] for the other two tasks, given its capability on accepting multi-modal inputs.

The single-task and multi-task models share the same model architecture as `Instruct-Imagen`, but **do not** have access to the *multi-modal instruction* during fine-tuning and inference. Instead, they accept the raw multi-modal inputs from each task. Additionally, the single-task model requires an independent model for each task, thereby inducing $7\times$ more parameters than `Instruct-Imagen`.

Human Evaluation. We follow the same evaluation protocol as [20] to conduct systematic human study. Each sample is rated by at least three raters for their semantic consistency score (SC) and perceptual quality score (PQ). The score in each category are $\{0, 0.5, 1\}$, where 0 means inconsistent / extremely poor quality and 1 means totally consistent / high quality respectively. Note that semantic consistency is defined as the score of the least consistent condition when there are multiple conditions. The final human score is defined as $O=\sqrt{SC\times PQ}$. We recruit eight huamn raters and train them following the guidelines[1] in ImagenHub [20]. Each method is evaluated independently, but we assign the same rater for samples generated by different methods given the same input to ensure evaluation calibrated per example.

[1]https://imagenhub.readthedocs.io/en/latest/Guidelines/humaneval.html

5.2. Main Results

Figure 4 compares `Instruct-Imagen` with our baselines and prior methods, showing it achieves similar or superior results in terms of in-domain evaluation and zero-shot evaluation (the breakdown of SC and PQ is detailed in the appendix). It suggests that multi-modal instruction training enhances performance in tasks with limited training data, such as stylized generation, while maintaining effectiveness in data-rich tasks, such as photorealistic imaging. Without multi-modal instruction training, our multi-task baseline tends to yield inferior image quality and text alignment. For instance, in the in-context stylization example of the Figure 5, the multi-task baseline struggles to differentiate style from subject, and replicate the subject in its generation. For similar reason, it generates 0 performance in the task of style transfer. This observation underscores the value of instruction tuning.

Distinct from many current approaches that rely on task-specific methods (e.g., StyleDrop [42] + DreamBooth [37]) or training [21], `Instruct-Imagen` efficiently manages compositional tasks by merging instructions for individual tasks, and inference in-context (no fine-tuning, taking 18.2 seconds per example). As shown in Figure 6, `Instruct-Imagen` consistently outperforms others in instruction following and output quality. Furthermore, in the presence of multiple references in the multi-modal context, the multi-task model fails to correspond the text instructions to the references, resulting in the ignorance of some multi-modal conditions. These results (more in Appendix A.1) further demonstrate the efficacy of the proposed model.

5.3. Model Analysis & Ablation Study

Besides the main results, we also perform studies to explore the limit of `Instruct-Imagen`, ablate important design of its training, and analyze its failure mode.

Fine-tuned `Instruct-Imagen` can edit image. Aside from zero-shot compositional tasks, another advantage of

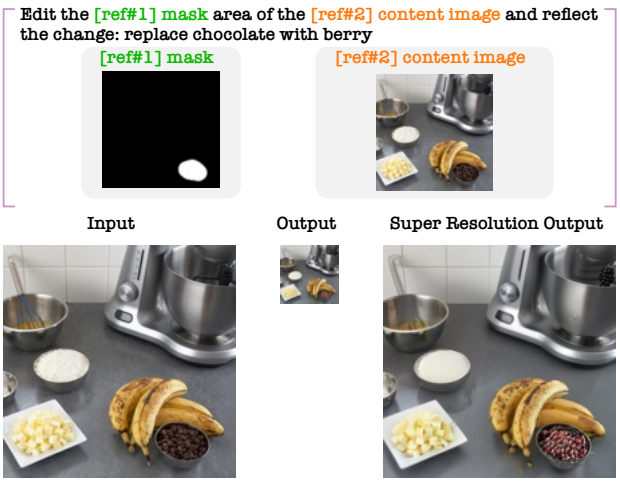


Figure 7. **`Instruct-Imagen` for masked image editing** When fine-tuned on MagicBrush [49], although `Instruct-Imagen` can edit the image as instructed (*i.e.*, see the 128×128 output), the super-resolution model fails to capture details from the input image, and causes the inconsistency.

Method	Setup	Human Score	Accuracy
`SDXL-inpainting`	-	0.43	0.25
`Imagen`	Fine-tuned	0.37	0.10
`Instruct-Imagen`	Fine-tuned	0.72 (+0.35)	0.57 (+0.47)

Table 2. Masked Image Editing Evaluation on ImagenHub [20].

`Instruct-Imagen` lies in its adaptability to new tasks. Particularly, we fine-tuned `Instruct-Imagen` on the MagicBrush dataset [49] ($\sim$ $9K$ examples) for $10K$ steps, and evaluated on the masked image editing data by ImagenHub [20]. We report the results using the overall score [20] (O), and the accuracy (*i.e.*, % of examples where SC=1). As a result, Table 2 presents a comparison between prior methods (SDXL-inpainting [29]), fine-tuned `Imagen` model (has been retrieval-augmented trained but without instruction tuning), and fine-tuned `Instruct-Imagen`. It shows that once fine-tuned, `Instruct-Imagen` can perform significantly better than the baseline method, and also method specifically designed for mask-based image editing. However, the fine-tuned `Instruct-Imagen` introduces artifacts into edited images, particularly in high-resolution outputs after super-resolution, as depicted in Figure 7. This occurs due to the model's lack of prior learning in pixel-accurate copying from context to output, a task significantly distinct from other `Instruct-Imagen` tasks.

Retrieval-augmented training helps generation quality. We compare variants of `Instruct-Imagen` in terms of whether performing retrieval augmented training and report results in Table 3. It shows the retrieval augmented training is a crucial step to obtain superior empirical results, in terms of both in-domain and zero-shot evaluation. This validates our hypothesis that retrieval augmented training benefits representing and handling multi-modal context.

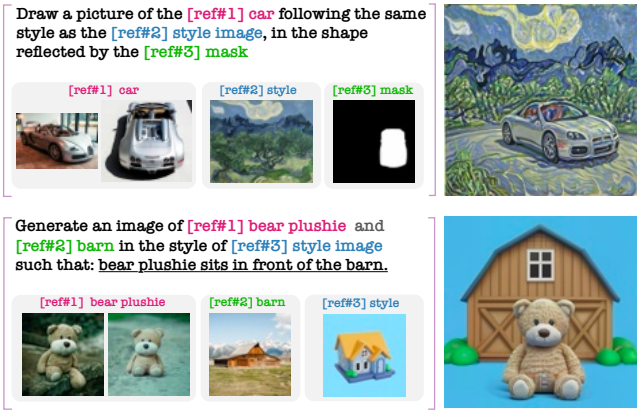


Figure 8. **Failure mode of `Instruct-Imagen`.** The most common failure of `Instruct-Imagen` is its incapability to follow each control condition in the instruction faithfully.

Method	In-domain Eval	Zero-shot Eval
w/o Retrieval-augmented	0.55	0.53
w/ Retrieval-augmented	0.79 (+0.25)	0.59 (+0.06)

Table 3. Ablation study on retrieval-augmented training. We report the average in-domain and zero-shot eval scores O.

Failure mode of `Instruct-Imagen`. One common pattern we found in `Instruct-Imagen` (when attempting complex instructions with $\geq$ 3 conditions) is its failure to follow instruction in the generation. Particularly, the model can accomplish the generation to satisfy only a subset of conditions specified in the multi-modal instruction. For instance, Figure 8 top example shows the model succeed to handle the style and subject, but do not generate the output in the shape that the mask specified.

6. Discussion

We introduce `Instruct-Imagen`, an image generation model that comprehends multi-modal instruction to accomplish a variety of visual generative tasks. It marks an initial but significant leap forward general-purpose visual generative model, via allowing not only in-domain image generation, but also zero-shot image generation on unseen and complex instructions. While opening up a new research direction, `Instruct-Imagen` can not handle image editing tasks in zero-shot. A key limitation is its lack of pixel consistency with input images, hindering the inclusion of additional tasks like in-painting and image editing in the instruction-tuning. This issue stems from the use of a cascaded diffusion model, which depends on a low-resolution model for crucial decisions like layout and object semantics. Such a low-resolution model struggles with both accessing high-resolution input details and reproducing them in the output, leading to artifacts in the generated image — because the super resolution model has to hallucinate the details. Based on this observation, we believe that one promising future direction is developing diffusion models that operate at the raw image resolution.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 4

[2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. PaLm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 2

[3] James Betker, Gabriel Goh, Li Jing, Brooks Tim, Jianfeng Wan, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, and Yunxin Jiao. Improving image generation with better captions. *Technical Report*, 2023. 5

[4] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. VisIT-Bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023. 5

[5] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 5

[6] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-Imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 2, 3, 4, 5

[7] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. In *NeurIPS*, 2023. 3, 4, 5, 7, 13

[8] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. PaLI: A jointly-scaled multilingual language-image model. In *ICLR*, 2022. 4, 5, 12, 13

[9] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. 4, 5

[10] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*, 2023. 5

[11] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2, 5

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 11

[13] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. In *BMVC*, 2017. 5, 7, 13

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 5

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

[16] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 2022. 3

[17] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. *arXiv preprint arXiv:2302.11154*, 2023. 5

[18] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 7

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 4, 5, 13

[20] Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. ImagenHub: Standardizing the evaluation of conditional image generation models. *arXiv preprint arXiv:2310.01596*, 2023. 2, 5, 7, 8

[21] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 5, 6, 7

[22] Dongxu Li, Junnan Li, and Steven CH Hoi. BLIP-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 3

[23] Mengtian Li, Zhe Lin, Radomir Mech, Ersin Yumer, and Deva Ramanan. Photo-Sketching: Inferring contour drawings from images. In *WACV*, 2019. 4, 5, 13

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 5

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (CelebA) dataset. *Retrieved August*, 2018. 4, 5, 13

[26] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 5

[27] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 5, 13

[28] 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. 7

[29] 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. 7, 8, 14

[30] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019. 5, 12

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

[32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 3

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 5

[34] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 5, 12

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

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

[37] 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. 2, 3, 5, 7

[38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2, 3, 4, 5, 11, 14

[39] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 11

[40] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *ICML*, 2018. 11

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

[42] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. StyleDrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 2, 3, 5, 7, 13

[43] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *TIP*, 2018. 4

[44] Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *TIP*, 2019. 5, 6, 12

[45] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 2, 5

[46] Saining "Xie and Zhuowen" Tu. Holistically-nested edge detection. In *ICCV*, 2015. 5, 12

[47] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 5

[48] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 5

[49] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. MagicBrush: A manually annotated dataset for instruction-guided image editing. *arXiv preprint arXiv:2306.10012*, 2023. 8

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

[51] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023. 5

[52] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023. 4