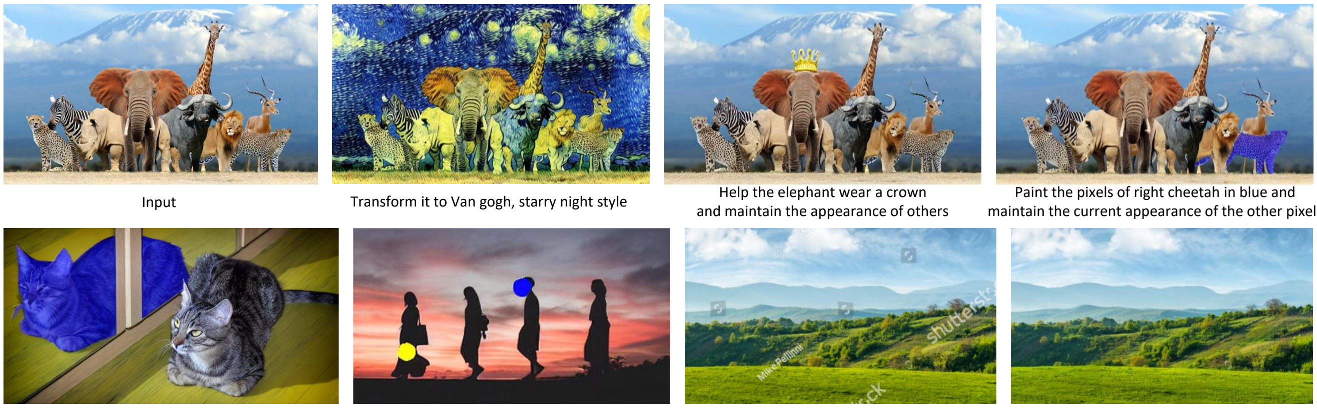
InstructDiffusion: A Generalist Modeling Interface for Vision Tasks

Zigang Geng[1,5*], Binxin Yang[1,5*], Tiankai Hang[2,5*], Chen Li[3,5*], Shuyang Gu[5†],
Ting Zhang[4], Jianmin Bao[5], Zheng Zhang[5], Houqiang Li[1], Han Hu[5], Dong Chen[5], Baining Guo[5]
[1]University of Science and Technology of China [2]Southeast University
[3]Xi'an Jiaotong University [4]Beijing Normal University [5]Microsoft Research Asia
https://gengzigang.github.io/instructdiffusion.github.io/



Input

Transform it to Van gogh, starry night style

Help the elephant wear a crown and maintain the appearance of others

Paint the pixels of right cheetah in blue and maintain the current appearance of the other pixels

Mark the pixels of cat in the mirror to blue and leave the rest unchanged

Use yellow to encircle the left knee of the people on the far left and draw a blue circle over the nose of the tallest people

Withdraw the watermark applied to this photograph

Figure 1. We introduce InstructDiffusion, a generalist modeling interface for vision tasks. Given input image and human instruction, our unified model effectively accomplishes tasks such as image editing, segmentation, keypoint estimation, and low-level vision.

Abstract

*We present InstructDiffusion, a unified and generic framework for aligning computer vision tasks with human instructions. Unlike existing approaches that integrate prior knowledge and pre-define the output space (*e.g.*, categories and coordinates) for each vision task, we cast diverse vision tasks into a human-intuitive image-manipulating process whose output space is a flexible and interactive pixel space. Concretely, the model is built upon the diffusion process and is trained to predict pixels according to user instructions, such as encircling the man's left shoulder in red or applying a blue mask to the left car. InstructDiffusion could handle a variety of vision tasks, including understanding tasks (such as segmentation and keypoint detection) and generative tasks (such as editing and enhancement) and outperforms prior methods on novel datasets. This represents a solid step towards a generalist modeling interface for vision tasks, advancing artificial general intelligence in the field of computer vision.*

1. Introduction

Recently, the field of artificial intelligence has witnessed remarkable advancements, particularly in natural language processing (NLP) [7, 13, 52, 53]. The Generative Pre-trained Transformer (GPT) has unified multiple NLP tasks using a single, coherent framework. Inspired by this, our research aims to achieve a similar unification in the realm of computer vision, *i.e.* [10, 11], developing a unified model capable of tackling multiple vision tasks simultaneously. However, unifying computer vision tasks is more challenging than NLP tasks due to the diversity of various tasks.

Diversity of Tasks and Outputs: Computer vision incorporates a wide range of tasks like segmentation, image generation, and keypoint detection. These tasks possess varying output formats, including masks, images, and coordinates. This diversity makes it challenging to find a uniform representation. In contrast, NLP tasks often have text-based outputs that can be easily represented in a standard format.

Different Methodologies and Techniques: Computer vision tasks often require unique methodologies and techniques depending on the specific problem. For example, image generation tasks are commonly dominated by Generative Adversarial Networks (GANs) [17, 28, 29] and Denois-

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

ing Diffusion Probabilistic Models (DDPM) [19, 23, 65], which are rarely used for image understanding tasks such as object recognition or image classification. Similarly, models designed for various understanding tasks are incapable of accomplishing image generative tasks. In contrast, NLP tasks tend to rely on a more consistent set of techniques, such as Transformer-based models [71], which can be applied across various NLP applications.

Continuous Input and Output: Computer vision tasks typically involve continuous input and output, such as coordinates or images. This continuous nature poses a challenge in devising a unified approach that can accurately handle such data. Techniques like Vector Quantized-Variational AutoEncoders (VQ-VAE) [58, 70] can discretize this data, but the process introduces quantization errors, leading to inaccuracies in the results. This issue is less prominent in NLP tasks, where the input and output data can be more easily discretized [7, 13, 71] into text tokens.

In this paper, we take advantage of the DDPM and propose a novel approach to address these challenges by treating a wide variety of tasks as image generation, specifically instructional image editing. We instruct image editing tasks using a more natural and intuitive way that closely aligns with the way humans process images. For instance, the instruction for a segmentation task could involve turning the pixels of an object in the image into a specific color, keeping the remaining pixels unchanged. Compared with previous methods that have attempted to formulate vision tasks as inpainting problems [5, 78], our approach accurately reflects human intentions, simplifying the handling of multiple vision tasks. At the same time, as the input and output of DDPM are continuous [23, 65], discretization is unnecessary, which solves the problem of quantization error.

We mainly focus on unifying three types of output formats: 3-channel RGB images, binary masks, and keypoints, which are sufficient to cover most vision tasks, such as semantic segmentation, referring segmentation, keypoint detection, image manipulation, and so on. Since the output of the denoising diffusion model is a 3-channel image, we propose a unified representation that encodes masks and keypoints into 3-channel images to handle various image understanding tasks. Then we use a post-processing module to extract the standard output format for evaluation.

During training, we use a diverse set of tasks to train a single, unified model. We also collect a new dataset for image editing. Our experimental results reveal that our model performs well across all tasks and outperforms previous methods on datasets unseen in training. Remarkably, we find that, compared to training individual models for each task, joint training of multiple tasks enhances the generalization. This study presents a solid step towards the development of a generalist modeling interface for vision tasks.

2. Related Work

The longstanding goal in artificial intelligence research is to develop a universal model that can solve any task. We present a brief overview of recent efforts towards this goal.

Vision Language Foundation Models. The vast amount of easily accessible web-scale image-text pairs has spurred innovative research in vision language foundation models [15, 35, 39, 45, 66, 75, 97]. The pioneering works, CLIP [57] and ALIGN [26], use contrastive loss to align image-text pairs in a shared cross-modal embedding space, demonstrating strong generalization. Subsequent efforts extend them to a broader spectrum, such as the image-text-label space proposed in UniCL [88] and a wider range of tasks as well as modalities supported in Florence [94] and INTERN [62]. However, these methods lack the ability to generate language, which limits their use in open-ended tasks like captioning or visual question answering.

Recently, the success of large language models like GPT series [7, 53, 55, 56], PaLM [4, 12], and LLaMA [68], has been attracting a lot of interest [24, 38, 64, 69, 73, 74, 79] in enhancing these models with visual capabilities. Mostly, these models cast a variety of open-ended vision tasks as text prediction problems, mapping visual input content to language semantics. BEIT3 [77] unifies the pretraining task in a masked modeling manner. CoCa [92] and BLIP [36, 37] unify contrastive and generative learning. Flamingo [2] accepts interleaved visual data and text as input and generates text in an open-ended manner. LLaVA [43] exploits visual instruction tuning by converting image-text pairs into an instruction-following format. GLIP v2 [95] and Kosmos v2 [54] leverage grounded image-text pairs to unlock the grounding capability of multimodal large language models. Our work sets itself apart by attempting to formulate vision tasks, such as segmentation and keypoint detection, into an instruction-following framework. This is challenging due to the unclear instructions and the diverse continuous output formats in these tasks.

Vision Generalist Models. The computer vision community aspires to develop a unified model capable of addressing a variety of tasks. Multi-task learning [25, 25, 33, 99] has gained popularity, but the diversity and complexity of task outputs present a key challenge. Currently, there are two major interfaces for output unification: language-like generation and image-resembling generation. Most existing attempts take inspiration from sequence-to-sequence models in the NLP field and model a sequence of discrete tokens through next token prediction [10, 20, 59, 74, 76]. Pix2Seq v2 [11] unifies object detection, instance segmentation, keypoint detection, and image captioning by quantizing the continuous image coordinates for the first three tasks. Unified IO [46] further unifies dense structure outputs such as segmentation masks and depth maps using a vector quantization variational auto-encoder (VQ-VAE) [70].

However, as quantization inevitably introduces information loss, another direction of unification explores the image itself as a natural interface for vision generalists [5, 78]. Painter [78] formulates the dense prediction task as a masked image inpainting problem, demonstrating in-context capability in various vision tasks. PromptDiffusion [80] also exploits in-context visual learning with a text-guided diffusion model [60], integrating the learning of six different tasks, i.e., image-to-depth, image-to-segmentation, and vice versa. Our work also examines image-resembling generation. Unlike previous works [78, 80] that derive the implicit task intention from in-context learning, our method introduces a more favorable instruction alignment. Moreover, with such explicit instructions, we further unify semantic image editing tasks, which are crucial use cases in image-resembling generation.

3. Method

We present InstructDiffusion, a novel generalist modeling interface for various vision tasks. Leveraging the Denoising Diffusion Probabilistic Model (DDPM), we treat all vision tasks as human-intuitive image manipulation processes with outputs in a flexible and interactive pixel space. Our primary focus is on three output formats: RGB images, binary masks, and keypoints, which effectively cover a broad spectrum of vision tasks, including keypoint detection, semantic segmentation, referring segmentation, semantic image editing, image deblurring, denoising, and watermark removal.

3.1. Unified Instructional for Vision Tasks

The unified modeling interface for all tasks is referred to as instructional image editing. By denoting the training set as $\{\boldsymbol{x}_i\}$, each training data $\boldsymbol{x}_i$ consists of $\{c_i, s_i, o_i\}$, where c_i is the control instruction, s_i and o_i represent the source and target images respectively. Our method aims to generate a target image o_i that adheres to the given instruction c_i when provided with an input source image s_i.

The semantic image editing dataset naturally consists of this type of triplet training data. For other vision tasks, the challenge lies in crafting suitable instructions and creating a corresponding target image. Although natural language instruction has been utilized widely in previous approaches, such as Pix2Seq [10] and UnifiedIO [46], we argue that terms like "segmentation" or "keypoint detection" in these methods are more likely to be treated by the model as indicators rather than instructions. To address this, we construct 10 diverse instruction templates for each task. During training, a template is selected randomly, turning the task's objective into detailed action instructions. These detailed instructions allow the model to fully understand the instructions rather than merely model a fixed bias based on the indicator. Next, we'll introduce example instructions for each task and the method for constructing the target image.

Keypoint detection. This task aims to accurately locate keypoint in an image, such as the left knee of an individual. For example, the instruction may be *"Use yellow to encircle the left knee of the people."* The edited image should show a yellow circle at the corresponding location while the rest of the region remains unaltered.

Segmentation. For semantic and referring segmentation, the goal is to identify the regions of specific objects within an image. An example instruction might be: *"Mark the pixels of the cat in the mirror to blue and leave the rest unchanged."* Consequently, the result is a edited image with a blue overlay marking the specified cat.

Image enhancement and image editing. Datasets for image editing, deblurring, denoising and watermark removal include both original and target images. Thus, we simply need to create instructions that precisely define the operation. For instance, *"Sharpen this image."* for deblurring, *"Remove the watermark."* for watermark removal, and *"Add an apple in the woman's hand."* for image editing.

3.2. Training Data Construction

As a proof-of-concept, we focus on investigating whether different tasks benefit each other under such image-resembling unification, instead of scaling data as much as possible for optimal performance at the extreme limits.

For image understanding tasks, we adopt widely used publicly available datasets and construct instructions and target images, as elaborated in Sec 3.1. For example, we use COCO-Stuff [8] for semantic segmentation. Further details about the datasets will be presented in Sec 4.1.

For semantic image editing, InstructPix2Pix (IP2P) employed a synthetic training dataset, using GPT-3 [7] for instruction generation and Prompt2Prompt [21] for output image creation. However, the resulted images show inconsistent quality and noticeable artifacts. MagicBrush [96] introduced a dataset of over 10,000 manually annotated triples, but its size is limited. Therefore, in addition to existing datasets such as IP2P [6], GIER [63], GQA [90], and MagicBrush [96], we propose a novel dataset called Image Editing in the Wild (IEIW), including 159,000 image editing pairs that cover a wide range of semantic entities and diverse levels of semantic granularity. To expand the scope of image editing data, we assemble the IEIW dataset from three distinct resources:

Object removal. Object removal is a common type of image editing. Inspired by Inst-Inpaint [90], we use the referring segmentation dataset PhraseCut [82] to build our instructional object removal data. PhraseCut offers images with referring phrases for corresponding regions. We set these regions as masks and use LAMA [67] to inpaint them, thus converting them into instructional inpainting datasets. Notably, we also swap input and output images, and reverse the instructions like "remove the blue bird on top of the tree"

Table 1. The number of effective training samples used for different tasks.

Task	Keypoint Detection	Segmentation	Image Enhancement	Image Editing
# Effective training samples	245k	239k	46k	425k

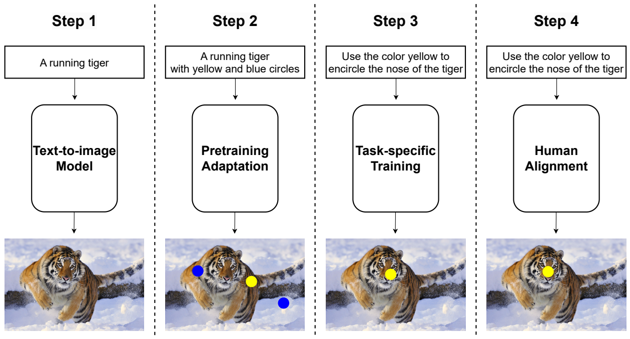

Step 1
Step 2
Step 3
Step 4
A running tiger
A running tiger with yellow and blue circles
Use the color yellow to encircle the nose of the tiger
Use the color yellow to encircle the nose of the tiger
Text-to-image Model
Pretraining Adaptation
Task-specific Training
Human Alignment

Figure 2. Training pipeline of our method. To illustrate concisely, we take keypoint detection as an example.

to "add a blue bird on top of the tree" to further supplement data from the perspective of adding components.

Object replacement. We propose a pipeline for generating training data for object substitution, an essential feature in image editing. Leveraging SA-1B [31] and OpenImages [32] datasets, which offer images with multiple semantically meaningful regions, we build a gallery database consisting of a variety of semantically-aware patches. Given a source image from OpenImages or SA-1B, we randomly select a semantic region and use it to retrieve its nearest neighbors from the gallery database. The source image and the retrieved similar patches are fed to PaintByExample [87] to generate a target image. We utilize an image captioning tool, such as BLIP2 [37], to create captions for the source and target images, and then generate possible instructions with a large language model. For example, given the captions "a running dog" and "a cute cat", a possible instruction is "please change the running dog to a cute cat". We can generate quite an amount of paired data for training using this construction pipeline.

Web crawl. To better align with user needs and enhance the user experience, we collect real-world Photoshop requests and their corresponding results from professionals via the websites. By using "photoshop request" as a keyword in Google, we gather over 23,000 relevant data triplets, which enhances our understanding of user requirements.

To ensure training data quality, we employ image quality assessment tools to eliminate substandard data. Specifically, we apply Aesthetics Score and GIQA [18] as evaluation metrics, utilizing LAION-Aesthetics-Predictor [61] for Aesthetics Score and constructing a KNN-GIQA model on LAION-600M [61] images for GIQA scores. Data is excluded if the target image has a low score or if there is a significant discrepancy between the source and its corresponding target image.

3.3. Unified Framework

Our framework is based on diffusion model, as it has demonstrated significant capabilities in modeling complex image distributions. As illustrated in Figure 2, our training procedure comprises three stages: pretraining adaptation, task-specific training, and human alignment.

Pretraining adaptation. Stable Diffusion [60], recognized as one of the most robust open-source text-to-image models currently accessible, serves as the foundation of our work, specifically utilizing version v1.5. While it maps textual captions to natural images, our desired images often involve segmentation masks or keypoint indicators, diverging significantly from typical natural images. Therefore, our preliminary stage aims to fine-tune the stable diffusion model and adjust the output distribution.

We augment the original image caption with a suffix, such as "with a few color patches." or "surrounded with circles." and add some random patches or circles on the original image as the target. By fine-tuning the diffusion model with these augmented text-image pairs, we enable the model to generate images within our desired output domain.

Task-specific training. In the second stage, we aim to fine-tune the diffusion model to enable it to accomplish various types of visual tasks. We follow InstructPix2Pix [6] and inject source images by concatenating them with the noise input, thus increasing the input channels of the first layer. We train our model using all data containing various tasks. Given the varying data volumes per task, we manually set different sampling weights for each database to ensure balance. Table 1 presents the number of effective training samples for each task. For each data triplet (c_i, s_i, o_i), the pretrained encoder $\mathcal{E}$ transforms the target image o_i into a latent representation z. The diffusion process then adds noise to z, resulting in a noisy latent representation z_t with the noise level increasing with each subsequent time step t. We finetune the diffusion network ϵ_θ by minimizing the following latent diffusion objective:

$$L = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),o_i,t,c_i,s_i}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_i, s_i)\|_2^2\right] \quad (1)$$

Human alignment. To further improve the quality of editing, we follow the idea of instruction tuning [81] from Large Language Models. While traditionally in LLMs, instruction tuning [81] trains models to solve a task following the instruction, our approach differs. We generate different editing results for each data sample using 20 distinct classifier-free guidance [22] parameters. We then invite subjects to select the best 0-2 edited images to formulate the instruction-tuning dataset, containing $1,000$ images. We use this dataset to fine-tune our model for about 10 epochs.

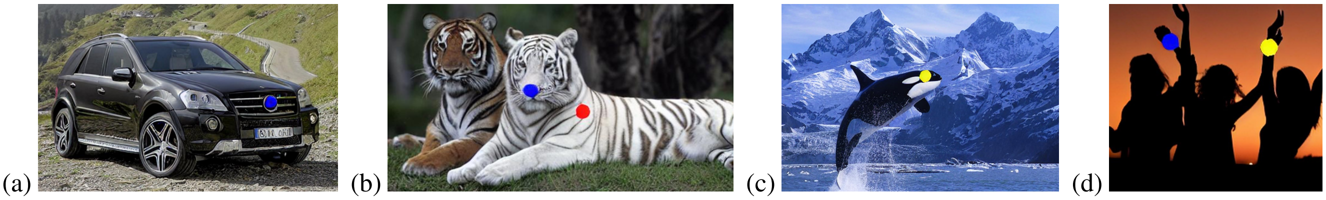


Figure 3. The keypoint detection results generated by our model. The instructions are as follows: (a) Mark the **car logo** with a **blue** circle. (b) Put a **blue** circle on the **nose** of the **white tiger** and use the **red** color to draw a circle around the **left shoulder** of the **white tiger**. (c) Create a **yellow** circle around the **right eye** of the **whale**. (d) Use the color **blue** to encircle the **right wrist** of the **person on the far left** and draw a **yellow** circle over the **left wrist** of the **person on the far right**.

4. Experiments

4.1. Settings

Training samples. Our model is trained on samples consisting of {instruction, source image, target image}, encompassing the aforementioned vision tasks, *i.e*., keypoint detection, semantic segmentation, referring segmentation, image enhancement, and image editing. For **keypoint detection**, we adopt four datasets: COCO [41], CrowdPose [34], MPII [3], and AIC [83], each containing 149K, 35K, 22K, 378K images respectively. During training, we randomly select 1-5 keypoints per image, assign them random colors, and generate the corresponding instruction. The target image is generated by positioning colored circles on the selected keypoints. For **segmentation**, we use COCO-Stuff [8] for semantic segmentation, gRefCOCO [42] and RefCOCO [93] for referring segmentation. We generate instructions from prompt templates, like "place a color mask on object". During training, we randomly assign a color for a category in semantic segmentation or an object in referring segmentation. For **image enhancement**, we utilize the GoPro [50] containing 2103 images, the REDS [51] dataset with 24,000 images for deblurring, the SIDD [1] dataset composed of 320 images for denoising, and the CLWD [44] dataset containing 60,000 images for watermark removal. Lastly for **image editing**, as mentioned in Sec. 3.2, we adopt 7 editing datasets, including filtered InstructPix2Pix [6] dataset containing 561K samples, 8K samples in MagicBrush [96] training dataset, GIER [63] with 5K samples, GQA [90] inpainting dataset with 131K samples, VGPhraseCut [82] composed of 85K samples, our generated dataset with 51K produced samples, and an internal dataset representing real editing scenario, which contains 23K training triplets.

Implementation details. We utilize Stable Diffusion v1.5 [60] for initialization and train our model with a fixed learning rate of 1×10^{-4}. The resolution of the input images is 256×256. The training involves a batch size of 3072 over 200 epochs, necessitating around four days on 48 NVIDIA V100 GPUs. We employ an EMA rate of 0.9999 during training and adjust it to 0.99 during the human alignment stage for rapid adaptation. Once trained, the model is instantly deployable for a variety of vision tasks.

Table 2. Average precision comparison on the COCO val2017, HumanArt and AP-10K datasets. We evaluate the official large models of the competitors to ensure fairness. The ground truth bounding boxes are used for all results. The best-performing generalist models are highlighted in bold.

Method	COCO val	HumanArt	AP-10K
Specialized Models			
PCT [16]	80.2	63.7	14.6
ViTPose [86]	82.0	64.1	14.7
Generalist Models			
Unified-IO [46]	25.0	15.7	7.6
Painter [78]	70.2	12.4	15.3
Ours	**71.2**	**51.4**	**15.9**

4.2. Keypoint Detection

We assess our model's performance using the COCO validation dataset, and its generalization ability on unseen datasets: HumanArt dataset [27] and AP-10K animal dataset [91]. The HumanArt, consisting of various forms such as cartoons, shadow plays, and murals, has a distinct data distribution from COCO. The AP-10K animal dataset demonstrates our model's capability to handle animal keypoints, despite being trained only on human ones. For a detailed evaluation, we use a lightweight U-Net structure to post-process the output images into the multi-channel heatmaps, thus extracting precise pose coordinates. We adopt the standard average precision based on OKS as our evaluation metrics. Specifically for the AP-10K dataset, OKS is calculated only on the keypoints overlapping with the COCO annotated joints for comparison with other methods. However, it should be noted that our model can detect keypoints beyond the confines of the training dataset.

Table 2 demonstrates that our model outperforms other general-purpose models, Unified-IO [46] and Painter [78], across all datasets. Notably, our superior performance on HumanArt and AP-10K underlines our framework's strong generalization ability. While our unified model does not surpass methods specifically tailored for keypoint detection due to localization accuracy limitations, it shines on the entirely unseen animal keypoints dataset, AP-10K. Figure 3 (a-c) display our model's capabilities in detecting car logos and animal keypoints that have never appeared in the keypoint detection training dataset. Figure 3 (d) further exhibits our model's versatility in referring keypoint detection.

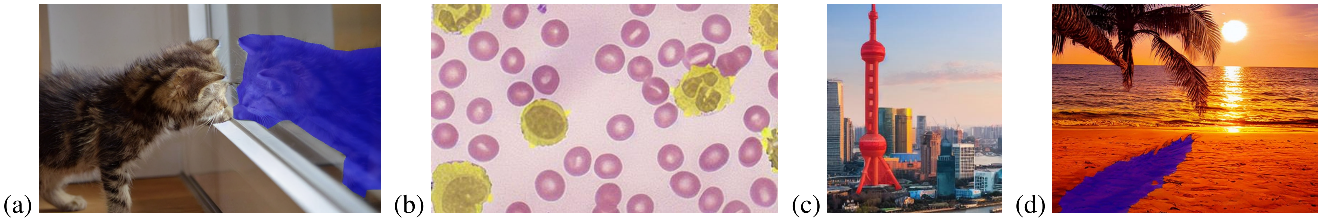


Figure 4. The segmentation results generated by our model. The instructions are as follows: (a) Mark the pixels of the **cat in the mirror** to **blue** and leave the rest unchanged. (b) Fill in the pixels of the **neutrophils** with **yellow**, retaining the existing colors of the remaining pixels. (c) Modify the pixels of **Oriental Pearl Tower** to **red** without affecting any other pixels. (d) Paint the pixels of the **shadow** in **blue** and maintain the current appearance of the other pixels.

Table 3. Referring segmentation results (cIoU↑). U: UMD split. G: Google split. The best-performing generalist models are bolded.

Method	gRefCOCO	RefCOCO			RefCOCO+			G-Ref			RefClef			
	val	val	test A	test B	val	test A	test B	$\text{val}_{(U)}$	$\text{test}_{(U)}$	$\text{val}_{(G)}$	val	testA	testB	testC
						Specialized Models								
LAVT [89]	57.64	72.73	75.82	68.79	56.86	62.29	48.14	58.65	59.17	61.16	21.22	44.77	24.78	47.08
ReLA [42]	62.42	73.21	75.24	68.72	56.10	62.26	47.89	59.71	60.40	61.37	20.68	43.08	22.57	45.94
						Generalist Models								
Unified-IO [46]	17.31	46.42	46.06	48.05	40.50	42.17	**40.15**	48.74	49.13	44.30	40.13	43.33	48.07	32.47
Ours	**67.36**	**61.74**	**65.20**	**60.17**	**46.57**	**52.32**	39.04	**51.17**	**52.02**	**52.18**	**49.58**	**54.73**	**54.82**	**40.34**

Table 4. Quantitative results on semantic segmentation (mcIoU↑). In the table, "A" stands for ADE20K (A-847, A-150), "P" stands for Pascal Context (P-459, P-59), "VOC" stands for Pascal VOC, and "CO" stands for COCO-stuff. The best-performing generalist models are highlighted in bold.

Method	A-847	P-459	A-150	P-59	VOC	CO
			Specialized Models			
SimSeg [40]	10.43	13.98	25.89	53.55	39.27	40.26
OvSeg [84]	13.85	22.72	36.50	60.93	38.50	48.76
SAN [85]	18.84	33.32	38.79	63.31	46.14	50.15
			Generalist Models			
Painter [78]	5.00	8.68	**54.50**	33.67	4.67	11.91
PromptDiffu. [80]	0.99	2.19	8.97	13.07	11.69	2.71
Unified-IO [46]	8.96	13.69	15.65	27.21	31.46	22.52
Ours	**19.68**	**28.29**	33.62	**59.00**	**72.55**	**53.17**

4.3. Segmentation

Our model is evaluated under two scenarios: the closed-set scenario using the COCO-stuff [8], gRefCOCO [42], RefCOCO [93] datasets and the open-set scenario involving eight datasets, *i.e.*, RefCOCO+ [93], G-Ref [47], RefClef [30] for referring segmentation, ADE20K-150 [98], ADE20K-847 [98], Pascal Context-59 [49], Pascal Context-459 [49], Pascal VOC [14] for semantic segmentation. We employ a U-Net structure to extract the binary mask of each object or each category. Following the previous method [42], we adopt cumulative IoU (cIoU) to measure the performance for referring segmentation. Perceiving semantic segmentation as referring segmentation based on semantic categories, we employ the mean of class-wise cumulative intersection over union (mcIoU) for its evaluation.

For referring segmentation, Table 3 showcases that our model significantly outperforms the generalist model Unified-IO [11] across almost all datasets, and even excels over models specifically designed for referring segmentation on the RefClef dataset. For semantic segmentation, Table 4 offers a quantitative comparison. In the closed-set scenario, *i.e.*, the COCO-Stuff dataset, our approach surpasses other models. In the open-set scenario, our method exceeds other models in both ADE20K-847 and VOC. For other datasets, we also significantly outperform other generalist models, with the exception of Painter for ADE20K-150, as Painter was trained on this dataset.

Intriguingly, we notice that both Painter [78] and PromptDiffusion [80] lack awareness of the colors associated with unseen categorie during open-set scenario evaluations. This limitation stems from their dependency on example images for color-semantics instruction. Conversely, our approach assigns colors to semantic categories via text instructions, resulting in significantly superior performance. Figure 4 illustrates several visual examples for segmentation to demonstrate our model's capability.

4.4. Image Enhancement

Our model's low-level vision performance is evaluated using widely used benchmarks, *i.e.*, GoPro [50], SIDD [1] and CLWD [44] for deblurring, denoising, and watermark removal task, respectively. We utilize the standard PSNR as the evaluation metric. Our model's deblurring capability is testing on the GoPro benchmark at a 1280×720 resolution, while SIDD and CLWD are evaluated at a 256×256 resolution, aligned with other works.

The results in Table 5 show that compared with other existing generalist models such as Painter [78], our model can handle various image editing and enhancement tasks. How-

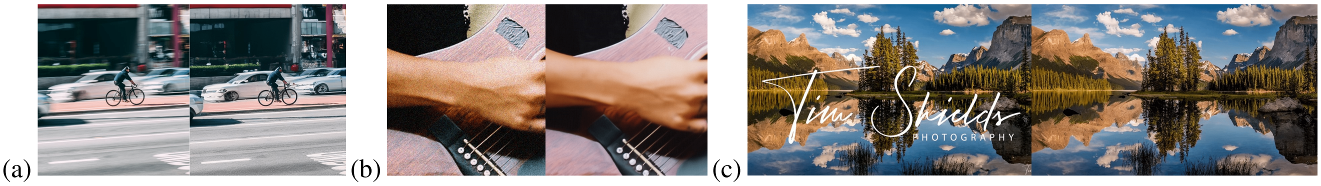


Figure 5. The results generated by our model for the low-level tasks. The instructions are as follows: (a) Sharpen this blurry image. (b) Purify this photo by removing noise. (c) Remove watermark from this picture.

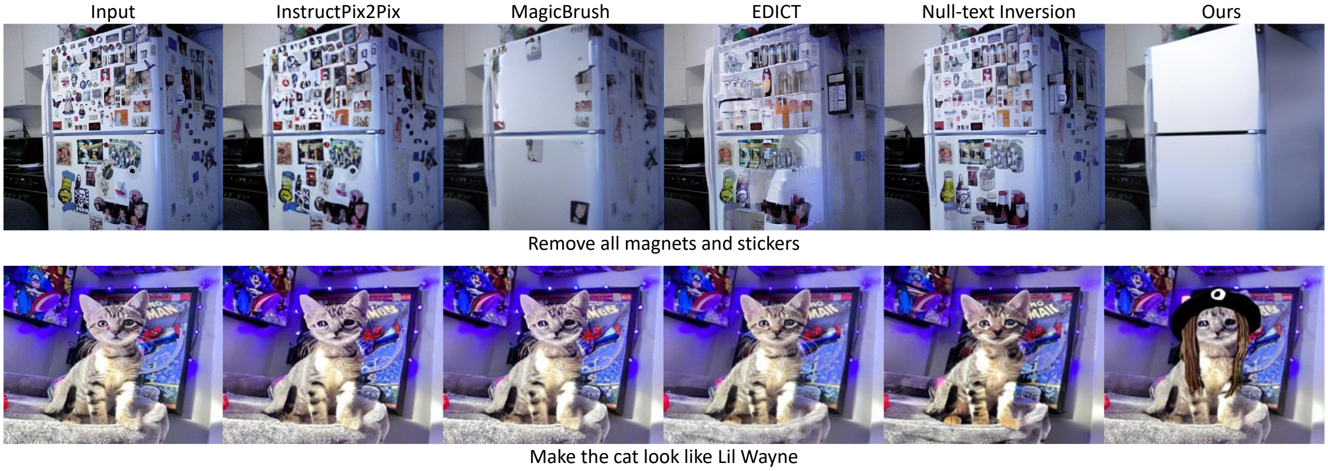

Input
InstructPix2Pix
MagicBrush
EDICT
Null-text Inversion
Ours
Remove all magnets and stickers
Make the cat look like Lil Wayne

Figure 6. Comparison between different instruction guided image editing. From left to right: input, Prompt-to-prompt [21], MagicBrush [96], EDICT [72], Null-text Inversion [48], and our results.

ever, the performance of our model in image enhancement is constrained by the VAE model, which introduces information loss. We have conducted an experiment by feeding the ground truth image to the VAE model and calculating the PSNR for the output, which serves as an upper bound for our model and is indicated in parentheses. Figure 5 displays our model's vision results in low-level vision tasks such as sharpening, denoising, and watermark removal, demonstrating its real-world effectiveness.

4.5. Image Editing

To showcase our method's editing quality, we created a benchmark of 1,000 samples. Each sample includes the source image, its BLIP2-provided [37] caption, the editing instruction, and the edited image's target caption. We manually sort samples into three distinct editing scenarios: replacement, removal, and addition, aiming to comprehensively reflect the model's editing capabilities. We adopt two commonly used metrics, CLIP similarity (CLIP-Sim) and Aesthetic Predictor's Score (AP) [61], to evaluate the editing results. CLIP-Sim measures the semantic similarity between an image and a text. We utilize BLIP2 [37] to obtain the caption of the input image and invoke GPT-3.5-turbo to acquire the target caption of the edited image. The CLIP-Sim score is calculated between the edited image and the target caption. The AP score assesses the aesthetic quality of the generated images, a methodology akin to LAION-5B, which employs the CLIP+MLP Aesthetic Score Predictor. A higher quality score reflects better perceptual quality.

The results in Table 5 show that our model achieves higher CLIP-Sim than InstructPix2Pix [6] and matches the result of MagicBrush [96]. The evaluation of the editing task must consider semantic resemblance, aesthetic appeal, background consistency, and manipulation accuracy. Quantitative comparison may be limited, as it may assign higher scores to models that make only minor edits, rather than those making substantial, meaningful changes.

Figure 6 illustrates visual examples compared with competitive methods including InstructPix2Pix [6], MagicBrush [96], EDICT [72] and Null-text Inversion [48]. Our model adeptly edits images based on provided instructions. More editing results are in the supplementary materials.

4.6. Ablation

Highly Detailed Instruction. We posit that generalization arises from learning through understanding the specific meanings of individual elements rather than memorizing entire instructions. Unlike earlier unified models like Pix2seq [10] and Unified-IO [46], which simply treat natural language as task indicators, our approach employs detailed task instructions, enabling the model to comprehensively understand and prioritize precise execution over mimicry. To show this, we substituted our instructions with simpler task indicators such as "semantic segmentation" and "keypoint detection", concurrently assigning fixed colors to each keypoint or object class. As shown in Table 6,

Table 5. Quantitative results on image editing and image enhancement. For editing tasks (Replace, remove, and add), the results are CLIP-Sim/AP score. For enhancement tasks, the number reflects the PSNR metric. The numbers in parentheses indicate the results obtained by reconstructing the ground truth images using VAE, representing the performance upper bound achievable with the used VAE model.

Method	Editing (CLIP-Sim↑ / AP Score↑)			LowLevel (PSNR↑)		
	Replace	Remove	Add	Deblur	Denoise	Watermark remove
Specialized Models						
NAFNet [9]	-	-	-	33.71	40.30	-
WDNet [44]	-	-	-	-	-	40.24
Null-text [48]	30.71/5.04	29.69/4.80	30.14/4.95	24.52	23.29	18.31
InstructPix2Pix [6]	29.61/4.99	28.82/4.69	30.11/4.94	22.71	15.14	14.96
MagicBrush [96]	30.50/4.94	29.07/4.67	30.69/4.90	22.64	16.10	15.46
EDICT [72]	29.91/4.91	29.33/4.80	30.19/4.93	24.16	24.48	19.88
Generalist Models						
Painter [78]	-	-	-	-	38.66	-
Ours	30.19/4.90	28.88/4.65	30.39/4.87	23.58 (29.54)	34.26 (36.56)	23.60 (24.80)

Table 6. Ablation study on the instruction. The term "Simple Instruction" denotes coarse-grained instructions like "semantic segmentation" or "keypoint detection". Our approach utilizes highly detailed instructions. COCO and COCO-stuff are included in the training set.

Method	Keypoint Detection			Semantic Segmentation					
	COCO	HumanArt	AP-10K	COCO-Stuff	ADE-847	PC-459	ADE-150	PC-59	VOC
Simple Instruction	22.7	7.0	5.2	41.28	1.39	3.96	13.65	18.49	20.22
Ours	71.2	51.4	15.9	53.17	19.68	28.29	33.62	59.00	72.55

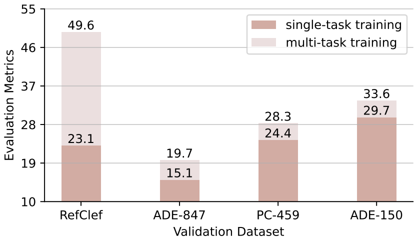


Figure 7. Ablation study on multi-task training. We evaluate our models on four unseen datasets, RefClef, ADE20K-847, PC-459, and ADE20K-150. It demonstrates that joint training significantly enhances the capability to handle open-set scenarios.

simple instructions yield notably inferior results, especially with novel keypoints or object categories. This highlights that our detailed instructions provide enhanced flexibility and adaptability in the open domain.

Multi-task Training. Multi-task learning, simultaneously addressing multiple tasks, often surpasses single-task learning in model generalization. Figure 7 validates this, comparing a model trained only on the segmentation dataset against our multi-task model over four unseen test datasets. The jointly trained model clearly performs better in open-domain testing scenarios. Furthermore, we observe that this benefit also extends to image editing. As demonstrated in Figure 8, the model, when integrated with other tasks, can more accurately identify objects needing editing, possibly due to the benefits of integrated referring segmentation.

5. Conclusion

In this work, we present InstructDiffusion, an innovative, unifying framework for aligning computer vision tasks with human instructions. InstructDiffusion treats various vision tasks as image generation, with a focus on three types of output formats: RGB images, binary masks, and keypoints. Our approach has demonstrated strong performance in each task and the generalization capability enhanced by joint training of multiple tasks. Notably, our method outperforms previous methods on unseen datasets. This research marks a solid step towards a generalist modeling interface for vision tasks and sets the stage for future advancements in the pursuit of artificial general intelligence in computer vision.

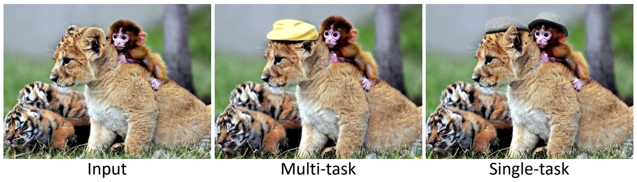


Figure 8. When given the instruction "put a hat on the leopard", joint training with other tasks can lead to more precise editing outcomes in comparison to single-task training.

In future work, we plan to focus on the following aspects to enhance the performance and capabilities of InstructDiffusion: 1) Broadening its application scope: We intend to incorporate a wider variety of tasks, including various task combinations, with the goal of showcasing its impressive potential for generalization. 2) Improving the unified representation: We aim to explore alternative encoding schemes and techniques to better represent a more diverse range of outputs associated with various computer vision tasks. 3) Investigating the role of self-supervised and unsupervised learning: we will explore the use of self-supervised and unsupervised learning techniques to leverage large-scale unlabeled data for model training and adaptation.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5, 6

[2] 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. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 5

[4] 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

[5] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022. 2, 3

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3, 4, 5, 7, 8

[7] 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. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2, 3

[8] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018. 3, 5, 6

[9] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022. 8

[10] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 1, 2, 3, 7

[11] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35:31333–31346, 2022. 1, 2, 6

[12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[14] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep*, 2007(1-45):5, 2012. 6

[15] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013. 2

[16] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human pose as compositional tokens. In *CVPR*, 2023. 5

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[18] Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Giqa: Generated image quality assessment. In *European conference on computer vision*, pages 369–385. Springer, 2020. 4

[19] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2

[20] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16399–16409, 2022. 2

[21] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022. 3, 7

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[24] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 2

[25] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 2

[26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[27] Xuan Ju, Ailing Zeng, Wang Jianan, Xu Qiang, and Zhang Lei. Human-art: A versatile human-centric dataset bridging

natural and artificial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[30] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 6

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 4

[32] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 4

[33] Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, et al. Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2691–2700, 2023. 2

[34] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, 2019. 5

[35] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[36] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 4, 7

[38] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[39] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 2

[40] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 6

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[42] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *CVPR*, 2023. 5, 6

[43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2

[44] Yang Liu, Zhen Zhu, and Xiang Bai. Wdnet: Watermark-decomposition network for visible watermark removal. In *2021 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2021. 5, 6, 8

[45] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[46] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 2, 3, 5, 6, 7

[47] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 6

[48] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 7, 8

[49] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 6

[50] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 5, 6

[51] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPR Workshops*, 2019. 5

[52] OpenAI. Gpt-4 technical report, 2023. 1

[53] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 1, 2

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

[55] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[56] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[57] 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. 2

[58] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2

[59] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022. 2

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3, 4, 5

[61] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 4, 7

[62] Jing Shao, Siyu Chen, Yangguang Li, Kun Wang, Zhenfei Yin, Yinan He, Jianing Teng, Qinghong Sun, Mengya Gao, Jihao Liu, et al. Intern: A new learning paradigm towards general vision. *arXiv preprint arXiv:2111.08687*, 2021. 2

[63] Jing Shi, Ning Xu, Trung Bui, Franck Dernoncourt, Zheng Wen, and Chenliang Xu. A benchmark and baseline for language-driven image editing. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3, 5

[64] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 2

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

[66] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[67] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022. 3

[68] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[69] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 2

[70] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[72] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22532–22541, 2023. 7, 8

[73] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems*, 35:5696–5710, 2022. 2

[74] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 2

[75] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016. 2

[76] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 2

[77] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a

foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[78] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 2, 3, 5, 6, 8

[79] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[80] Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. *arXiv preprint arXiv:2305.01115*, 2023. 3, 6

[81] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021. 4

[82] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020. 3, 5

[83] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: a large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017. 5

[84] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 6

[85] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 6

[86] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation, 2022. 5

[87] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 4

[88] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022. 2

[89] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022. 6

[90] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. Inst-inpaint: Instructing to remove objects with diffusion models. *arXiv preprint arXiv:2304.03246*, 2023. 3, 5

[91] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 5

[92] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[93] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 5, 6

[94] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[95] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022. 2

[96] 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. 3, 5, 7, 8

[97] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022. 2

[98] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 6

[99] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16804–16815, 2022. 2