Supplementary for UltraEdit: Instruction-based Fine-Grained Image Editing at Scale

Contents

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

Table 1: Datasets used to form the high-quality image caption dataset.

Dataset	#Samples	License	Annotator
MS COCO [15]	164,000	CC BY 4.0	Human
Flickr [32]	31,783	Custom	Human
NoCaps [1]	45,000	CC BY 2.0	Human
VizWiz Caption [3]	23,431	CC BY 4.0	Human
TextCaps [29]	28,408	CC BY 4.0	Human
Localized Narratives [21]	849,000	CC BY 4.0	Human
ShareGPT4V [5]	1,200,000	CC BY-NC 4.0	GPT-4V
LAION-LVIS [26]	220,000	Apache-2.0	GPT-4V

A Implementation Details

A.1 Collection of High-Quality Image Caption Data

To ensure the diversity and quality of the image editing pairs in ULTRAEDIT, our data generation pipeline relies on high-quality oracle data to mitigate biases within the generation models. Consequently, we focus on building the image editing dataset based on real images and their captions, enhancing the dataset's reliability in real-world scenarios and providing more comprehensive guidance for data generation than text-only data.

Like previous works [13, 5, 33, 36], we gathered source data from various public data sources with image caption data, as illustrated in Table 1, which includes diverse images with either manually annotated captions or detailed captions generated by advanced image captioning models [5, 26]. After filtering out images with excessively long or short captions, our collection amounted to 1.6 million high-quality image-caption pairs, which will be used for edit instruction generation and subsequent image generation.

A.2 Instruction and Caption Generation

To obtain high-quality editing instructions, we introduce a pipeline that combines human raters and language models for generating edit instructions and corresponding captions for subsequent image generation. Large language models have demonstrated remarkable abilities in various areas, such as agents [7, 11, 28, 14] and tool uage [25]. In our practice, we utilize the LLM to generate suitable image editing instructions. Firstly, we use language models to expand manually crafted edit examples to a set of 100,000 examples, as shown in Table 2. These examples serve as in-context learning examples to help the language models grasp the understanding of editing styles and requirements, enabling them to generate suitable editing instructions and corresponding edited caption. The query prompt is illustrated in Table 3.

We sample 50 editing instructions and 10 edit examples as in-context learning examples for querying the language model. The language model then generates appropriate editing instructions and corresponding edited captions for the given image captions from the collection. Leveraging the in-context-learning capabilities and generalization ability of the language model, we ultimately generate 4.16 million text-only data comprising creative yet sensible edit instructions and corresponding captions. Each case consists of a high-quality image caption for a real image, an editing instruction, and an edited caption corresponding to a target image.

A.3 Region-based Data Generation

For generating region-based image editing data, we first employ the recognize-anything [35] to identify objects in the source images. We then query the language model with the obtained object lists, edit instructions, and corresponding image captions to determine the target objects of the editing instruction using the query prompt in Table 4. If the editing instruction is object-oriented, the language model identifies the objects involved in the editing; otherwise, the entire image is considered as the editing target.

Table 2: Examples of instructions and their corresponding captions after Expansion.

Original Caption	**Edit Instruction**	**Edited Caption**
Two Asian dolls with big noses, fancy purple dresses, and golden hats.	Replace the dolls with miniature elephants in colorful traditional Indian cloth	Two miniature elephants wearing colorful traditional Indian cloth.
A man is watching TV in bed with only his foot showing.	Remove the bed and place the man in a deserted beach scene.	A man is watching TV on a deserted island beach with only his foot showing.
a person throwing a red frisbee, which is currently in mid-air, slightly to the right and above the person's hand.	Change the color of the frisbee to blue	a person throwing a blue frisbee, which is currently in mid-air, slightly to the right and above the person's hand.
A slipper near the edge of a concrete floor near small rocks.	Transform the slipper into a glass one	A glass slipper near the edge of a concrete floor near small rocks.
A woman wearing a shirt for the Religious Coalition for Reproductive Choice.	Change the background to a bustling cityscape at night	A woman wearing a shirt for the Religious Coalition for Reproductive Choice in front of a bustling cityscape at night.
A pot and some trays are in a kitchen.	Add a warm, inviting atmosphere to the image	The warm glow highlights a pot and some trays in a cozy kitchen.
a person that is jumping his skateboard doing a trick.	Turn the skateboard into a flying carpet	a person that is jumping his flying carpet doing a trick.
A stuffed bear is hanging on a fence.	Make it a snowy winter landscape	A stuffed bear is hanging on a snowy winter landscape fence.
A police dog wearing his bullet proof vest.	replace the background with a city skyline	A police dog wearing his bullet proof vest in a city skyline.
A baseball game is going with children playing a runner is about to hit the base.	Turn the baseball field into a magical forest	Children playing a runner is about to hit the base in a magical forest.
A small horse carries a women in a sled.	Turn the horse and sled into a spaceship traveling through outer space	A futuristic spaceship travels through outer space.
a person wearing a life jacket participating in water sports like water skiing.	Add a family of dolphins swimming around the person in the water	a person wearing a life jacket participating in water sports like water skiing, with a family of dolphins swimming around him in the water.

To generate region-based edited images, we use the Grounding DINO [16] to obtain bounding boxes of editing areas in real images, serving as coarse-grained masks. Subsequently, we perform SAM [12] on these bounding boxes to derive fine-grained object masks, expanding them to create contour masks that define the editing region. During image generation, we have observed irregular color boundaries between the editing region and the rest of the image. To ensure smooth transitions and high image quality, we fuse the fine-grained and bounding box masks to create a soft mask guiding the image generation. Specifically, for the editing region latent M_f and bounding box latent M_b, we fuse the two masks, making the region between them a soft mask region M_s. The generation pipeline can be formulated as follows:

$$z_{t-1} = \begin{cases} (1 - M_s) \cdot z_T + M_s \cdot DM(z_t) & \text{if } t \mod 2 == 0 \\ DM(z_t) & \text{otherwise} \end{cases} \tag{1}$$

Table 3: Query prompts for LLMs to write edit instructions and corresponding captions.

Prompt for Writing Edit Instruction	
Element	**Content**
`Intro`	`I will present a series of image editing instruction examples essential for mastering and understanding a variety of editing styles and requirements. Here are some sample instructions:`
`Instruction Examples`	`{instruction_str}`
`Task Description`	`I will provide one image caption corresponding to a specific image. You are required to apply the learned editing techniques to form suitable, detailed, and accurate editing instructions for the image defined by the caption. Note that your editing instructions should be distinct from the examples provided. Then, produce a description corresponding to the revised image after applying the editing instruction. Only necessary amendments should be made for the new image caption.`
`Output Format`	`The output format should be "original image caption; edit instruction; new image caption". Maintain the given format for the result. Please ensure to deliver solely the result, without incorporating any additional titles.`
`Image Caption`	`The image caption is: {caption}`
`Produce Instances`	`Produce three suitable instances based on the caption and return the list.`
`Examples`	`Here are some output examples for your reference: {example_str}`
`Response`	`Response:`

Additionally, we define M_s as:

$$M_s = \begin{cases} s & \text{for elements in } M_b \setminus M_f \\ M_f & \text{otherwise} \end{cases} \tag{2}$$

where M_f is the editing region latent, M_b is the bounding box latent, and s is the hyperparameter that determines the inpainting rate. During the generation, During the generation, we set s to range from 0.2 to 0.8.

A.4 An Improved Baseline for Free-form and Region-based Image Editing

We fine-tune the Stable Diffusion 1.5 model [23] using the Diffusers library [31] with data from UltraEdit. We maintain the hyperparameters as set in Brooks et al. [4]. Specifically, we train the model on 8 × 80GB NVIDIA A100 GPUs with a total batch size of 256. Following prior works [4, 33], we use an image resolution of 256 × 256 for training and 512 × 512 for generation.

To incorporate additional guidance from region masks, we concatenate the latent of the Region Mask M_s with the noisy latent Z_T and the latent of the source image Z_I to form the input to the diffusion model. We add four additional channels to the UNet of the diffusion model to accommodate the latent of the region mask M_s. The weights of the UNet are initialized with the pretrained diffusion model, while the extra eight channels (four for the source image latent Z_I and four for the mask latent M_s) in the convolutional layers of the diffusion UNet are randomly initialized. The model is then trained using a mixture of free-form and region-based image editing data from UltraEdit. For

Table 4: Query prompts for LLMs to capture objects that need editing.

Prompt for Capturing Editing Object	
Element	**Content**
Intro	The following prompt provides an instruction for image editing, an original image caption, a revised image caption that reflects the given edit instruction, and a set of objects detected by an object detection algorithm.
Edit Instruction	Edit Instruction: "{edit_instruction}"
Original Caption	Original Image Caption: "{input_text}"
Revised Caption	Revised Image Caption: "{output_text}"
Object List	Set of Objects Identified by the Recognition Model: {object_list}
Task Description	Your task is to identify the objects most likely to be modified based on the information provided above. Consider this from a comprehensive perspective; note that some objects might not be explicitly mentioned in the instructions or the Identified Objects list, but their appearance could still be affected. Please use precise words or phrases in your response.
Note	Note: 1. If you can't identify any specific edited object (e.g., a style transfer involving the entire image instead of a single object; add/move an object, which does not fit object-oriented editing instructions), please respond with "NONE". 2. Your response should exclusively identify the objects requiring edits, excluding any extra context or details. 3. Please list the objects to be edited, separating each one with a comma. The number of objects identified in the answer should not exceed 2.
Response	Response:

free-form image editing data, the model takes a blank mask as input to implicitly indicate that the editing should affect the entire image.

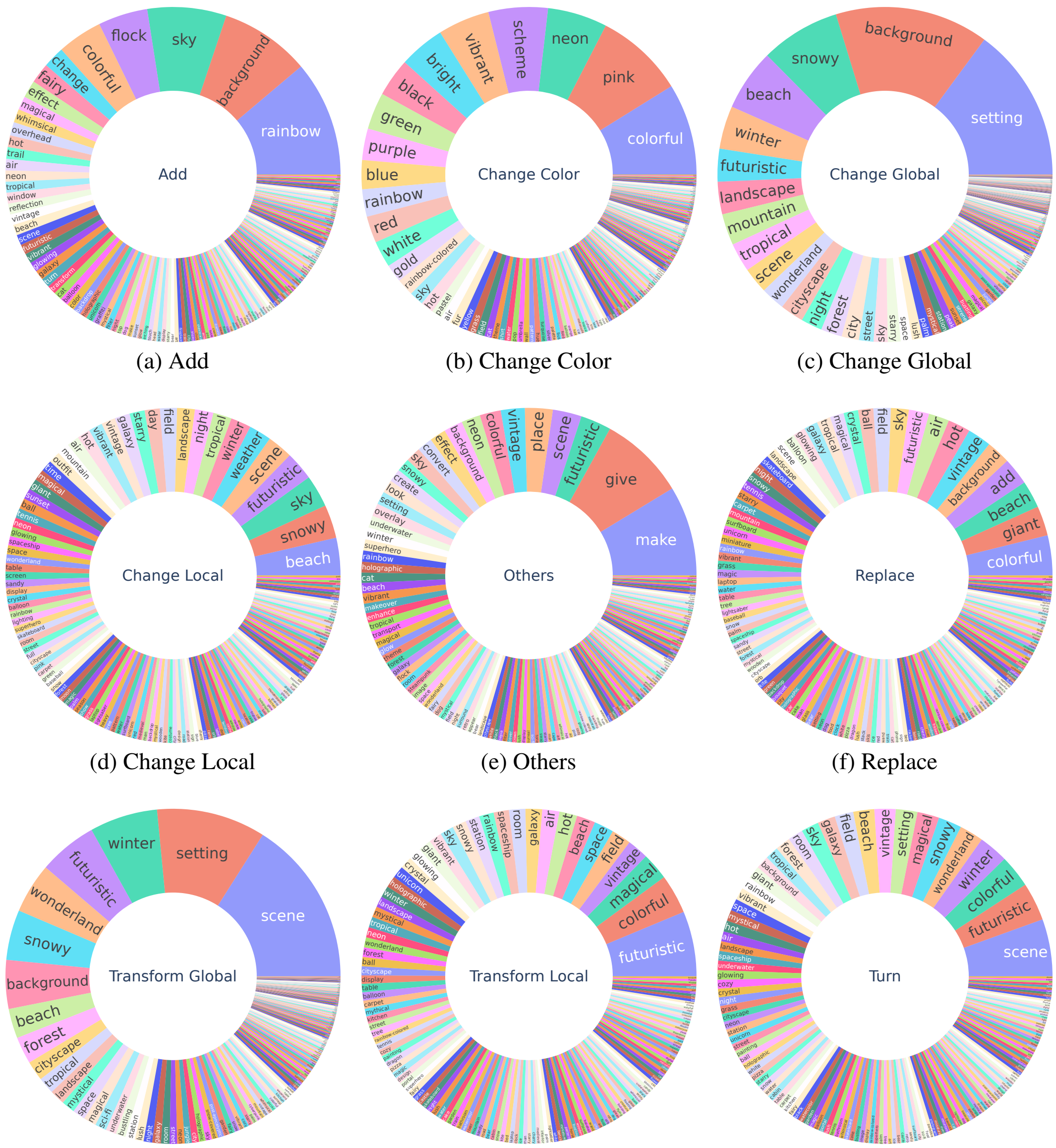
When training the model exclusively with Free-form Image Editing data, we strictly follow the settings of Brooks et al. [4] without making any additional modifications.

B Statistics of ULTRAEDIT

Table 5: Statistics of Free-form and Region-based Image Editing Data. The table shows the instance numbers, number of unique instructions, and their respective proportions for different instruction types.

Data Type	**Statistic**	**Change**			**Transform**		**Add**	**Replace**	**Turn**	**Others**	**Total**
		Color	**Global**	**Local**	**Global**	**Local**					
Free-form	**Inst. No.**	111,563	204,294	500,108	150,851	597,165	909,065	683,529	490,219	353,289	4,000,083
	Proportion (%)	2.79	5.11	12.50	3.77	14.93	22.73	17.09	12.26	8.83	/
	Unique Inst.	27,436	24,020	92,891	26,587	117,063	114,647	133,222	102,280	86,180	724,326
	Proportion (%)	3.79	3.32	12.82	3.67	16.16	15.83	18.39	14.12	11.90	/
Region-based	**Inst. No.**	2,912	3,515	15,774	2,796	21,807	11,918	25,749	16,628	7,080	108,179
	Proportion (%)	2.69	3.25	14.58	2.58	20.16	11.02	23.80	15.37	6.54	/
	Unique Inst.	1,056	727	5,032	762	6,835	3,201	8,064	5,256	2,620	33,553
	Proportion (%)	3.15	2.17	15.00	2.27	20.37	9.54	24.03	15.66	7.81	/

In this section, we dive into the characteristics and statistics of ULTRAEDIT. We present ULTRAEDIT, a large-scale, diverse, and high-quality real-image-based image editing dataset designed to advance the capabilities of image editing models. ULTRAEDIT comprises over 4,000,000 instruction-based free-form image editing instances and 100,000 region-based image editing instances, making it the largest open-source image editing dataset. Notably, it is also the first large-scale dataset focused on region-based image editing. Table 5 illustrates the statistics of the image editing data in ULTRAEDIT. It shows the numbers and proportions of the different instruction types and data types of ULTRAEDIT. Moreover, the Figure 1 shows the distribution of keywords in the instructions of ULTRAEDIT for various instruction types.



(a) Add (b) Change Color (c) Change Global

(d) Change Local (e) Others (f) Replace

(g) Transform Global (h) Transform Local (i) Turn

Figure 1: Distribution of keywords in the instructions of ULTRAEDIT for various instruction types

B.1 Comparison with other dataset

In this section, we compare our dataset with the InstructPix2Pix (IP2P) dataset in Free-form image editing. We report the results of automatic metrics to evaluate the data quality of each dataset.

As illustrated in Table 6, our dataset outperforms the InstructPix2Pix (IP2P) dataset across all tasks. Notably, the higher CLIPimg scores observed in some tasks for the IP2P dataset suggest that while the image pairs in the dataset may exhibit semantic similarity, it does not achieve the desired visual similarity, which is crucial for successful image editing. This highlights a fundamental shortcoming in the IP2P dataset and underscores the superior of our dataset.

Table 6: Comparison of dataset quality between ULTRAEDIT and InstructPix2Pix (IP2P) using automatic metrics. We evaluates the data quality across different tasks types.

Task	Dataset	CLIPin	CLIPout	CLIPdir	CLIPimg	SSIM	DINOV2
Change	Ours	0.2849	0.3024	0.2967	0.8441	0.6360	0.7403
	IP2P	0.2667	0.2661	0.2317	0.8557	0.5685	0.6194
Transform	Ours	0.2851	0.3005	0.2902	0.8289	0.6251	0.6875
	IP2P	0.2646	0.2680	0.1974	0.8667	0.5853	0.6972
Turn	Ours	0.2846	0.3018	0.2922	0.8321	0.6255	0.6949
	IP2P	0.2654	0.2698	0.2015	0.8575	0.5526	0.6419
Add	Ours	0.2786	0.3145	0.2957	0.8661	0.6645	0.7758
	IP2P	0.2629	0.2744	0.1990	0.8851	0.6318	0.7026
Others	Ours	0.2843	0.3038	0.2981	0.8374	0.6420	0.7048
	IP2P	0.2657	0.2706	0.1929	0.8629	0.5734	0.6847
Overall	Ours	0.2834	0.3049	0.2950	0.8427	0.6401	0.7231
	IP2P	0.2650	0.2694	0.1982	0.8660	0.5826	0.6859

B.2 More Examples of ULTRAEDIT

In this section, we showcase additional examples from ULTRAEDIT to illustrate the versatility and robustness of our dataset in various image editing tasks. The free-form editing data is depicted in the left two columns, while the region-based image editing data examples are in the right column. The examples highlight both Free-form and Region-based editing capabilities. It can be noticed that, due to using real images as anchors, our data shows high diversity in real-world scenarios, including text, natural environments, human figures, abstract objects, and even blurred low-quality images.

In Figure 2 and Figure 3, editing examples not only contain text modification, and abstract object editing, but also multi-step editing within a single instruction and fine-grained editing. Moreover, because of high-quality captions derived from open-source image caption datasets for generating editing instructions, the generated instructions are highly related to the source image. The region-based image editing data demonstrates high image element preservation in the editing examples. For instance, in the examples in the right column, the target images only perform edits within the masked area and keep the rest unchanged, even for highly blurred texts and human facial expressions in the figure.

C Baseline and Metrics

C.1 Baselines.

We set the following models as baselines, categorized into instruction-based image editing methods and global description-guided image editing methods, the latter requiring global descriptions of the target image to perform zero-shot editing. The instruction-based image editing methods include: InstructPix2Pix [4], HIVE [34], MagicBrush [33], and Emu Edit [27]. The global description-guided image editing methods include: Null Text Inversion [18], SD-SDEdit [17], GLIDE [19], and Blended Diffusion [2]. Notably, GLIDE and Blended Diffusion require a mask for editing.

Instruction-Based Methods:

- **InstructPix2Pix** uses automatically generated instruction-based image editing data to fine-tuning Stable Diffusion [24] and performance image editing based on the instructions during the inference, without any test-time tuning.

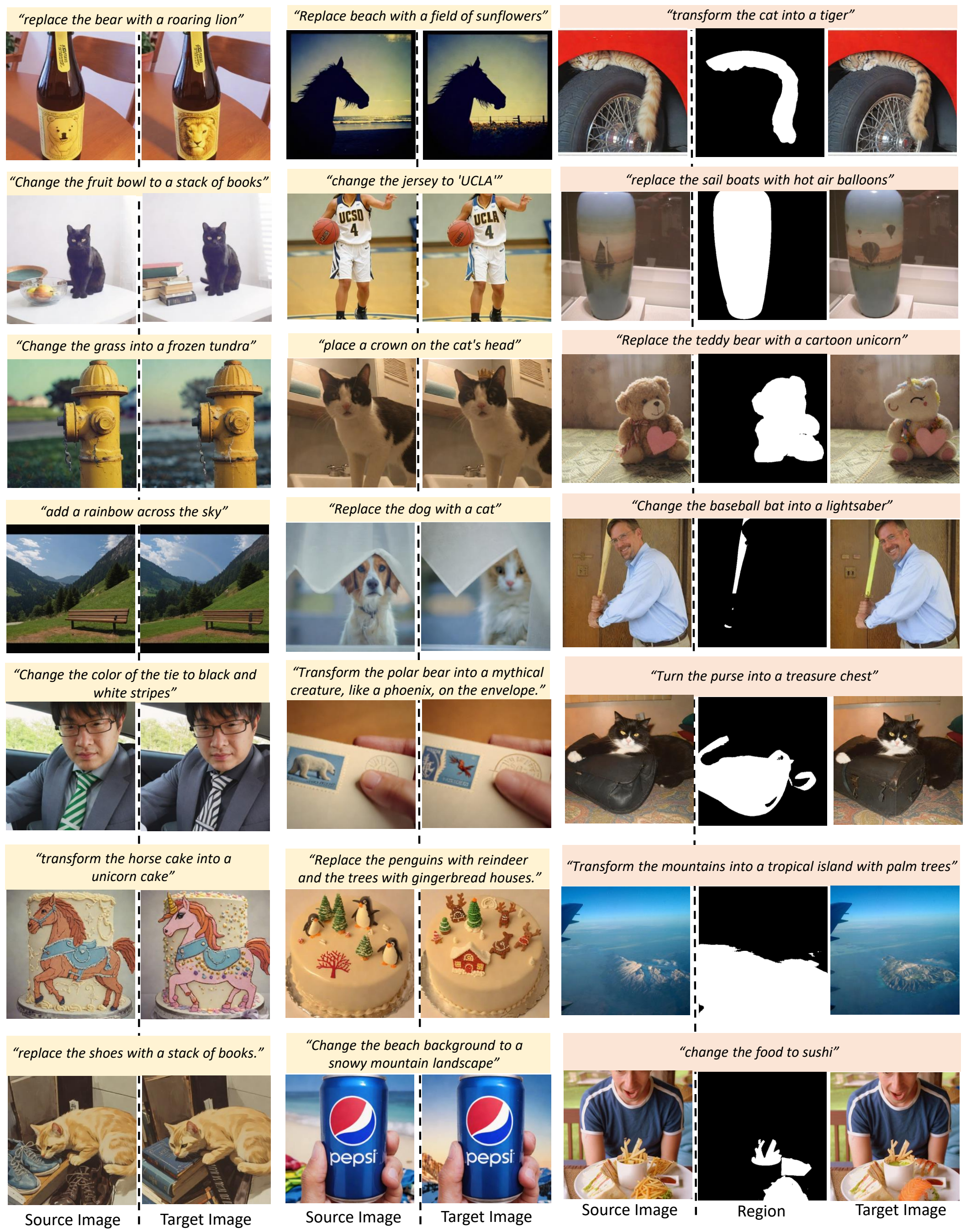


Figure 2: **More Examples of UltraEdit**. Free-form (left) and region-based (right) image editing.

- **HIVE** is trained with more data similarly to InstructPix2Pix, and is further fine-tuned with a reward model trained on human-ranked data.
- **MagicBrush:** is a variant of InstructPix2Pix, which is fine-tuned on the human-annotated dataset, MagicBrush.
- **Emu Edit:** is a closed-source model that supports multi-task image editing and achieves state-of-the-art performance. It is trained on a diverse set of tasks using 10 million of training data, including image editing and computer vision tasks.

Global Description-Guided Methods:

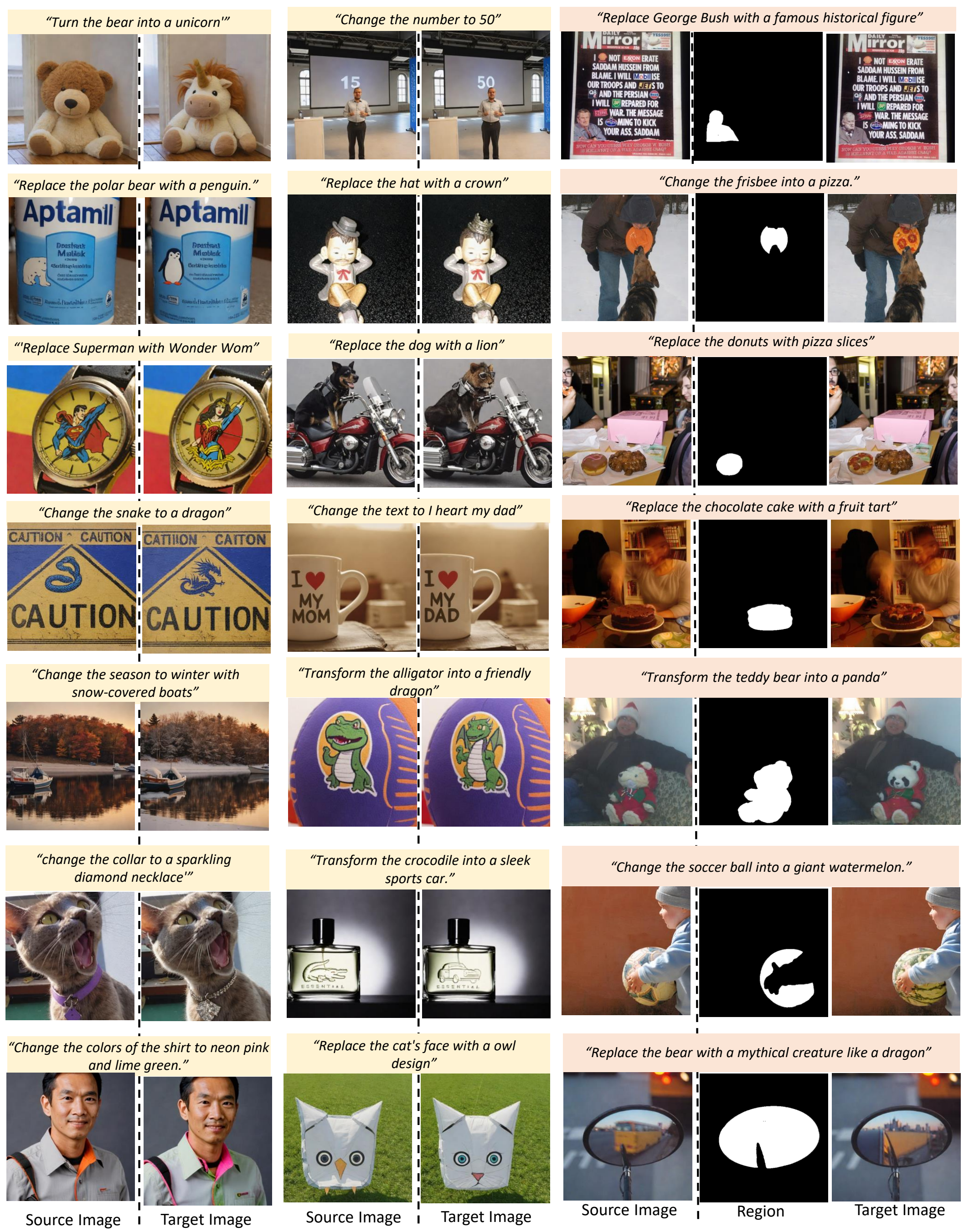


Figure 3: **More Examples of UltraEdit**. Free-form (left) and region-based (right) image editing.

- **Null Text Inversion:** inverts the source image with DDIM [30] trajectory and then performs editing during the denoising process with text-image cross-attention control [10].
- **SD-SDEdit:** noises the guidance image to an intermediate diffusion step, and then denoises it using the target description.
- **GLIDE:** is trained with 67M text-image pairs to fill in the masked region of an image conditioned on the local description with CLIP guidance.
- **Blended Diffusion:** blends the input image in the unmasked regions with the context in the noisy source image during each denoising timestep to enhance region-context consistency.

C.2 Details on Benchmarks and Metrics

MagicBrush aims for evaluating the single and multi-turn image editing ability of the model. It provides annotator defined instructions and editing masks, as well as the ground truth images generated by DALLE-2 [22] for evaluation, allowing for more effective metric assessment of the model's editing performance. However, this dataset also suffers from inherent bias. During data collection, annotators were directed to use the DALLE-2 image editing platform to generate the edited images. Thus, this benchmark is biased towards generated images and editing instructions that the DALLE-2 editor can successfully follow, which may compromise both its diversity and complexity. Following the setting of MagicBrush [33], we utilize L1 and L2 to measure the pixel-level difference between the generated image and ground truth image. And also adopts the CLIP similarity and DINO similarity to measure the overall similarity with the ground truth. Finally, the CLIP-T is used to measure the text-image alignment between local descriptions and generated images CLIP embedding.

Emu Edit Test aims for reducing bias of the annotator defined dataset and reach higher diversity. It contains the devise relevant, creative, and challenging instructions and high quality captions that capture both important elements in the image for source and target images, without any ground truth images. Consequently, consistent with the Emu Edit [27], we utilize the L1 distance, CLIP image similarity and DINO similarity between the **source images** and **edited images** to measure the the model‘s ability of preserving elements from the source image. Also, we use the CLIP text-image similarity between edited image and output caption and the CLIP text-image direction similarity(CLIPdir) to measure the instruction following ability of the model. Specifically, the CLIPdir measures agreement between change in caption embedding and the change in image embedding.Since the Emu Edit [27] does not specify the versions of the CLIP and DINO models used for the metric, we adopted the settings utilized by MagicBrush to maintain alignment with other benchmarks. Specifically, the versions are `ViT-B/32` for CLIP and `dino_vits16` for DINO embeddings. We ensure consistency by rerunning all results of different methods on Emu Edit benchmark. Additionally, there are known issues with the quality of the benchmark, wherein some image-caption pairs appear incorrect. These issues include placeholder captions (e.g., 'a train station in city') or instances where source and target captions are identical. To address these problems, we simply remove the incorrect cases prior to evaluation. Despite the Emu Edit Test eliminating bias and overfitting at the image level by not providing ground-truth images, the evaluation metrics still implicitly measure the model's editing ability.

D Qualitative and Human Evaluations

D.1 Human Evaluation

We conducted human evaluations to assess the consistency, instruction alignment, and image quality of the edited images generated by our model trained on ULTRAEDIT using the MagicBrush benchmark and Emu test benchmark. We first compared the performance of our model with the MagicBrush [33] and instructPix2Pix [4] models through a comprehensive human evaluation on MagicBrush benchmark. Additionally, we compared the performance of various models trained using our dataset with Emu Edit [27] on the Emu test benchmark. For the two evaluations, we randomly sampled 500 examples from the test sets of the MagicBrush benchmark and the Emu test benchmark, respectively.

For each sample, the evaluators compared the consistency, instruction alignment, and image quality of the edited images generated by the different models. As shwon in Figure 4, the evaluators were asked to determine which edited image was better by selecting between "First Image", "Second Image", or "Tie". The results are evaluated with TrueSkill [9] rating system. The scores of these evaluations are presented in Table 7 and Table 8. Our model (finetuned with our own ULTRAEDIT) can produce more preferable editing results than the baselines, even better than the MagicBrush baseline, which is reported to overfit on its test set.

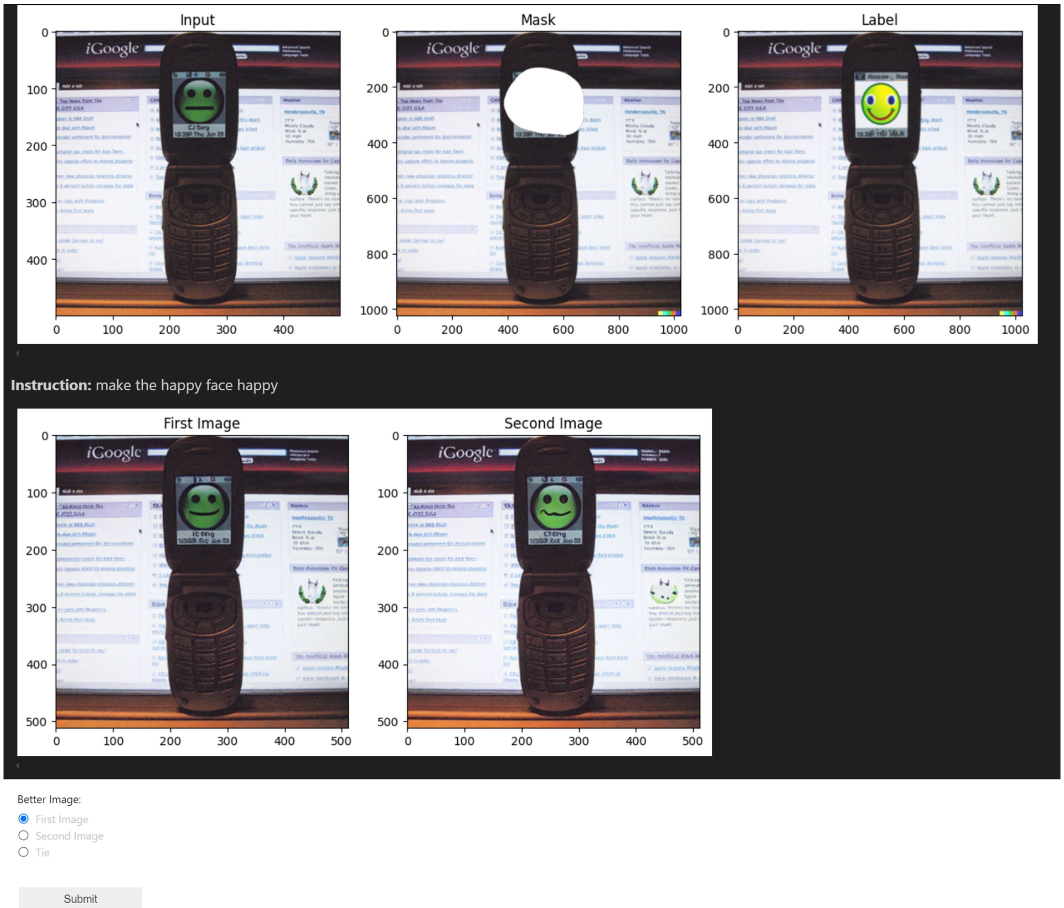


Figure 4: The interface of human evaluation on MagicBrush [33] and Emu test [27] benchmark to evaluate generated images by different models.

	Ours w/ UltraEdit	MagicBrush [33]	InstructPix2Pix [4]
TrueSkill score	25.5 ± 0.8	23.7 ± 0.7	22.6 ± 0.7

Table 7: TrueSkill [9] scores of image editing models evaluated by human raters on MagicBrush [33] test set.

D.2 Qualitative Evaluation on Different Benchmarks

In Figure 5 and Figure 6, we present the qualitative examples of different editing tasks on single-turn and multi-turn editing on MagicBrush. In Figure 7, we present the qualitative examples on Emu Edit Test across various editing tasks.

D.3 Qualitative Evaluation on Real Image Anchors

In Figure 8, we present qualitative results comparing the image editing generation method with and without using real images as anchors. Using real images as anchors to guide the data generation significantly enhances the diversity of the generated images and ensures that the generation results are more stable and aligned with the editing instructions. The image anchors provide substantial information for generation that goes beyond what is conveyed by the image captions alone. Specifically, image anchor ensures visual consistency between the generated source and target images in the image editing pairs, as shown in the first three rows of Figure 8. It can also be observed that with real

	SD3 [6] w/ UltraEdit	SDXL [20] w/ UltraEdit	SD1.5 [24] w/ UltraEdit	Emu Edit [27]
TrueSkill score	26.7 ± 0.7	26.5 ± 0.7	26.0 ± 0.7	25.1 ± 0.7

Table 8: TrueSkill [9] scores of image editing models evaluated by human raters on Emu Test [27] test set.

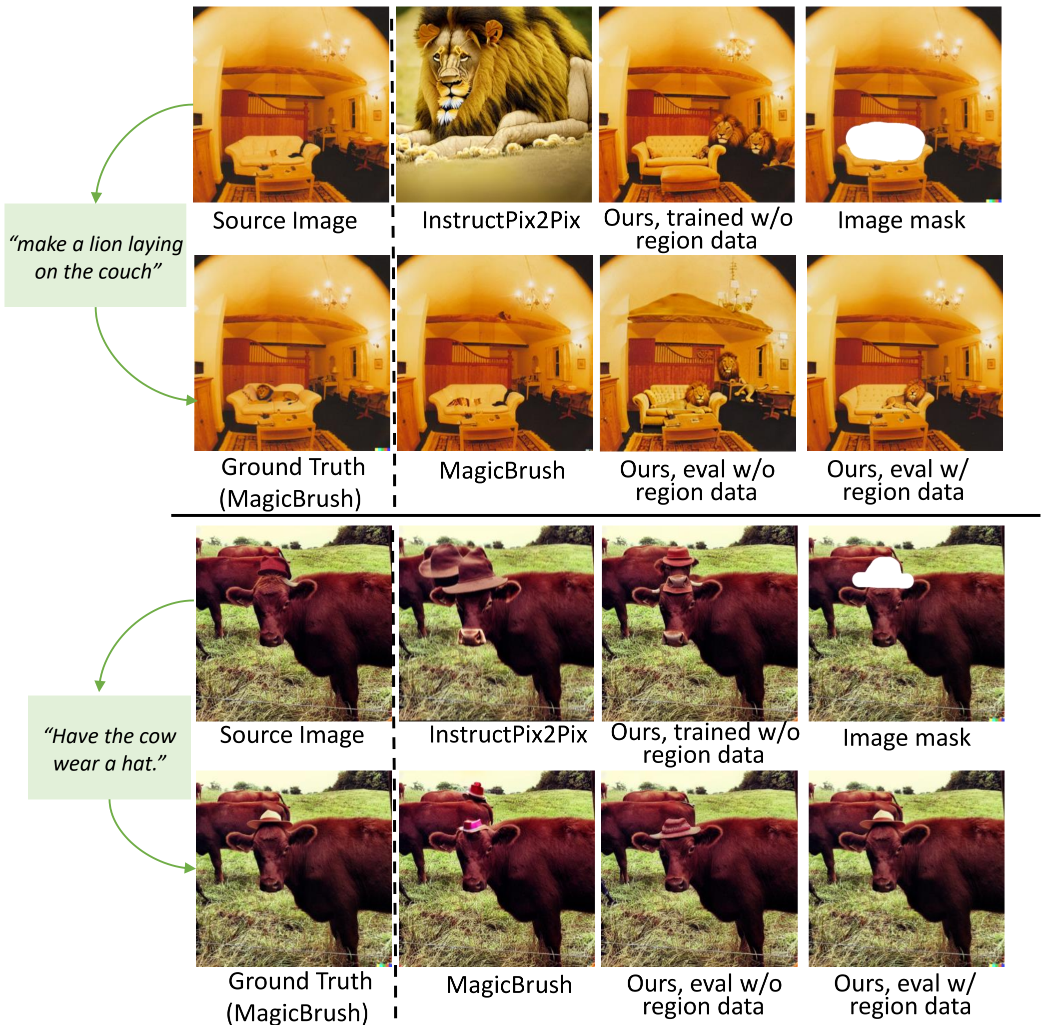


Figure 5: Qualitative evaluation of the model trained on UltraEdit across MagrichBrush benchmark in the Single-turn Setting.

image anchors, the editing process is more controlled, resulting in fine-grained image edits in the generated samples (see the last three rows in Figure 8).

D.4 Qualitative Evaluation on Free-form vs. Region-based Editing

In Figure 9, we present qualitative results comparing the model trained with an additional region-based editing task against the model trained solely with free-form image editing data. The comparison highlights that the inclusion of the region-based editing task during training enables the model to perform significantly more precise operations even in the absence of region input during evaluation, especially for background and localized edits.

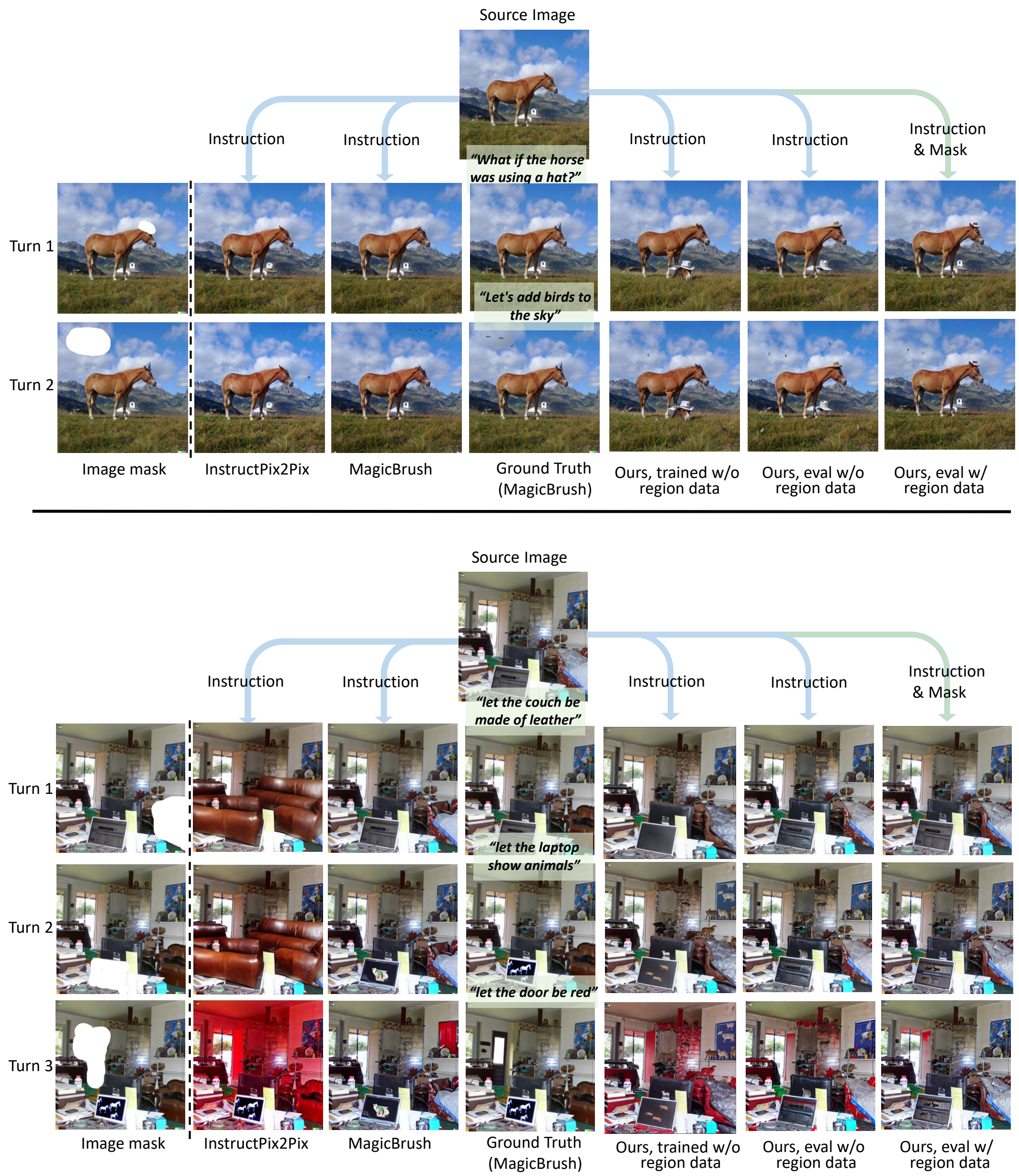


Figure 6: Qualitative evaluation of the model trained on ULTRAEDIT across MagrichBrush benchmark in the Multi-turn Setting.

D.5 Details of the region-based image editing pipeline

In Stage III, we apply our proposed method for generating region-based images to ensure a seamless transition between inpainted areas and the rest of the image. Initially, we analyze inaccurate masks generated by the segmentation model, as shown in Figure 11. We find these inaccuracies generally fall into a few categories: incorrect identification resulting in overly large masks, masks that are too small for effective editing, fragmented masks from segmentation failures, and fine-grained segment masks that closely resemble the original object, complicating the editing process.

To address these issues, we filter out excessively large, small, or fragmented masks. Fine-grained masks are adjusted using a soft mask, either a bounding box or contour mask. Our data circulation indicates that our methods significantly reduce artifacts and abrupt boundaries between the mask

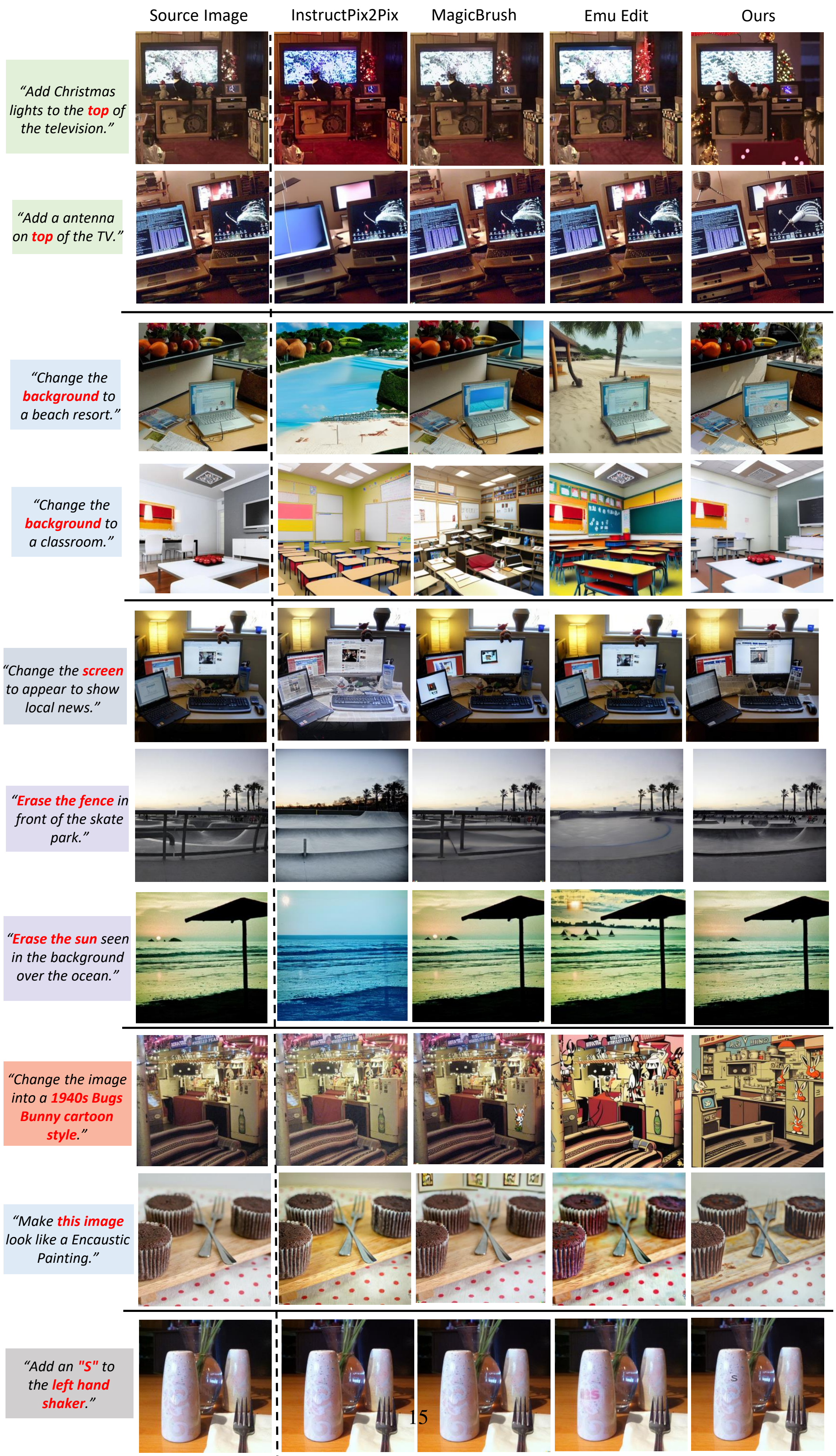


Figure 7: Qualitative evaluation of the model trained with ULTRAEDIT on the Emu Test.

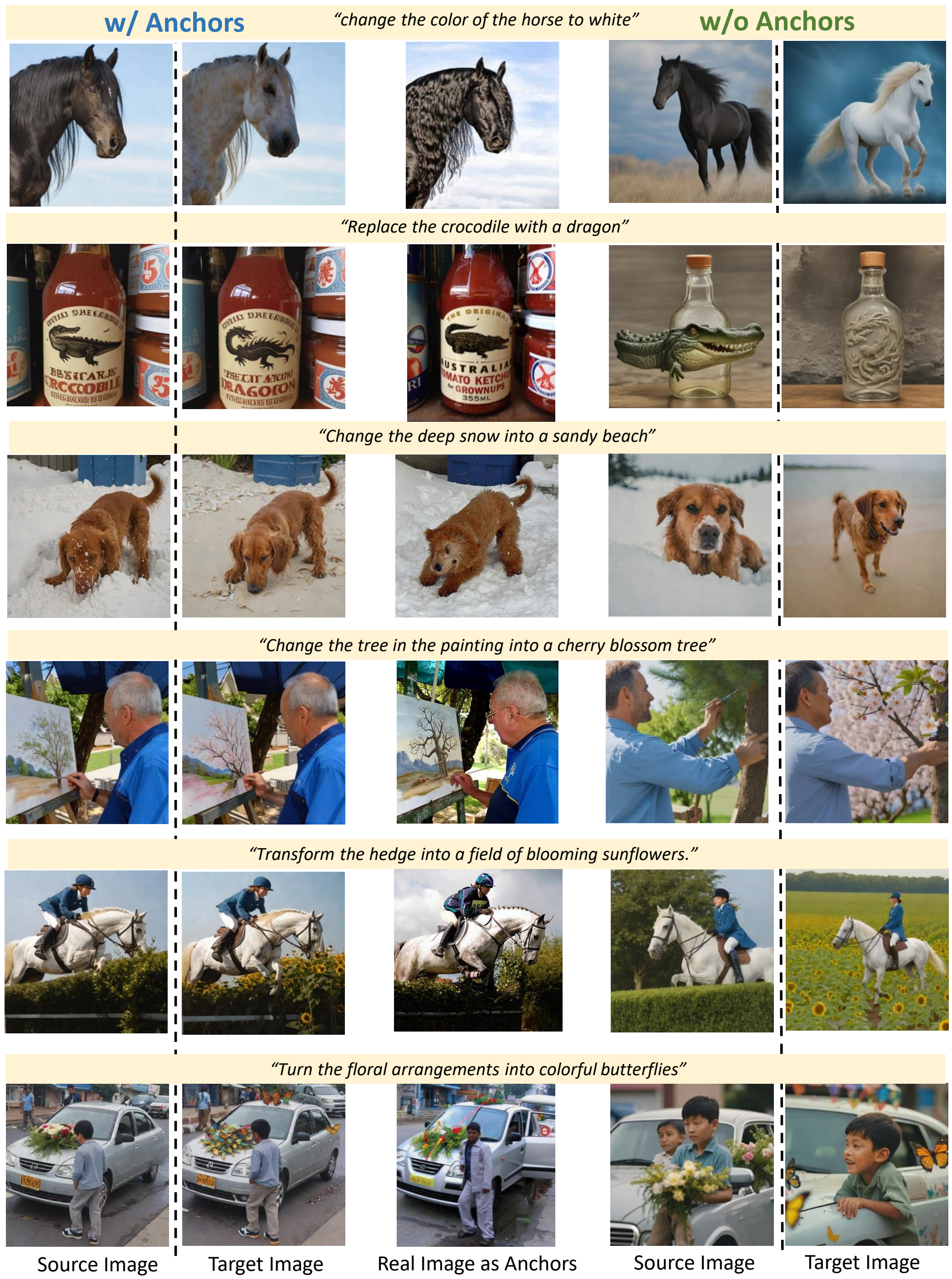


Figure 8: Qualitative evaluation of using real images as anchors during image generation. We compare qualitative examples between the generation pipeline using real image anchors (**left**) and the generation pipeline without real image anchors (**right**). The real images are presented in the middle column.

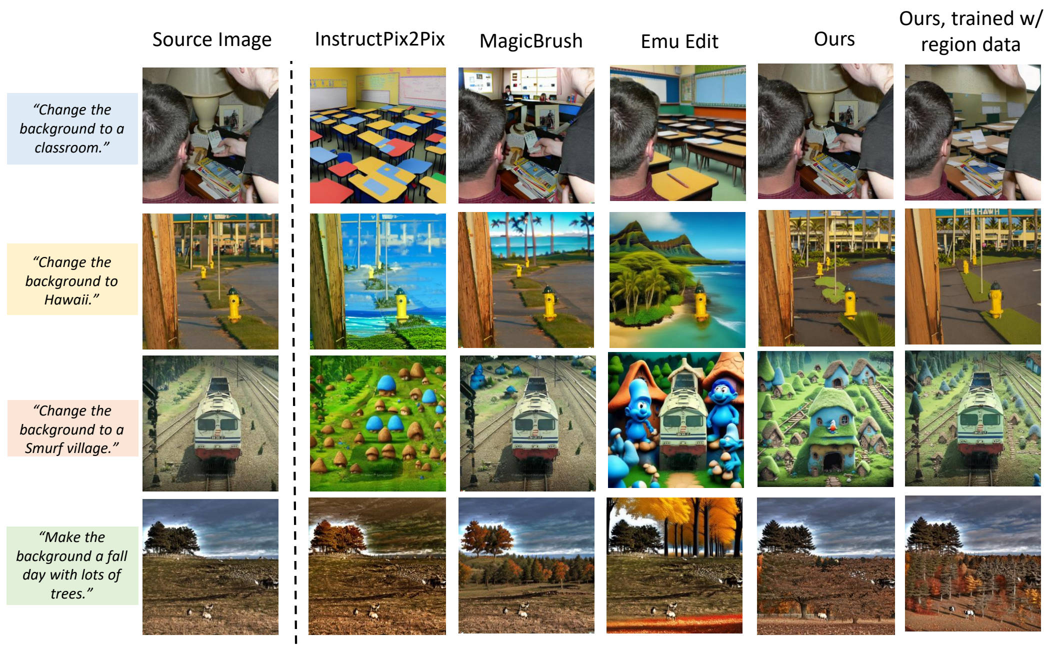


Figure 9: Qualitative evaluation comparing free-form and region-based editing task.

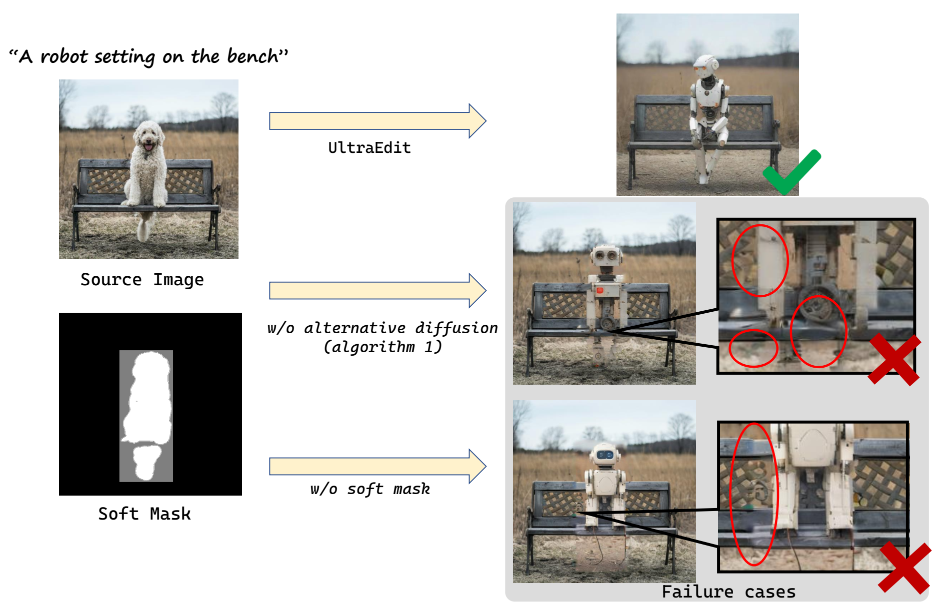
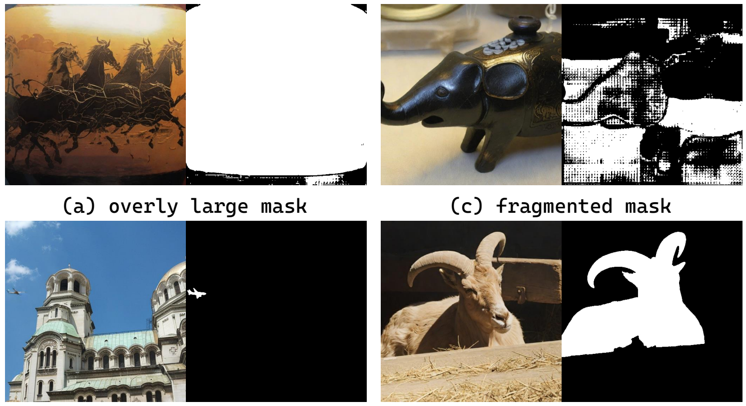


Figure 10: Qualitative evaluations of the region-based image editing pipeline. Generated images Without our method exhibit noticeable artifacts along the boundaries of the original and edited regions, emphasizing pronounced border effects.



(a) overly large mask (c) fragmented mask

(b) overly small mask (d) fine-grained mask

Figure 11: The four main categories of inaccuracies generated masks.

region and the remaining image. Qualitative evaluations shown in Figure 10 demonstrate the effectiveness of our approach. Images generated without our method exhibit noticeable artifacts along the boundaries of the original and edited regions, highlighting the advantages of our method.

E Statement on Limitations and Ethical Concerns

E.1 Limitations

While ULTRAEDIT represents a significant advancement in instruction-based image editing, several limitations should be acknowledged. Firstly, although our dataset includes diverse editing instructions and real image anchors, the reliance on automatically generated data may introduce some biases and errors. The quality and relevance of the editing instructions, influenced by current large language models and human raters, may not capture all nuances of creative and artistic editing tasks. Furthermore, despite our efforts to provide high-quality region annotations, there may be occasional inaccuracies or inconsistencies in the automatically produced region-based editing data.

Moreover, while our experiments demonstrate the benefits of using real image anchors and region-based editing data, the improvements shown by our diffusion-based editing baselines are benchmark-specific. They may not generalize across all editing scenarios. Future work should focus on enhancing the precision of region annotations and validating the dataset's applicability across a broader range of editing tasks.

Despite these limitations, ULTRAEDIT offers a robust and diverse dataset that significantly contributes to the field of image editing, paving the way for future research and development.

E.2 Ethical concerns

While UltraEdit offers substantial advancements in the field of instruction-based image editing, several ethical concerns must be considered:

- **Bias and Fairness.** The dataset, while diverse thanks to the efforts on real image anchors, *etc*., may still contain biases introduced by the automatic generation process and the inherent biases present in the large language models and human raters used. These biases could perpetuate stereotypes or unfair representations in the edited images.
- **Misinformation and Misuse.** The powerful image editing capabilities enabled by UltraEdit could be misused to create misleading or deceptive content, contributing to the spread of misinformation. It is crucial to implement safeguards and promote responsible use of the technology to mitigate this risk.
- **Privacy.** Real image anchors included in the dataset may contain identifiable information. Although efforts have been made to anonymize and protect personal data, there remains a risk of unintentional breaches of privacy.

To address these ethical concerns, we encourage users of UltraEdit to adhere to ethical guidelines, implement robust checks for bias and fairness, and prioritize transparency and accountability in their work. Additionally, we recommend ongoing dialogue within the research community to continuously refine and improve ethical standards in developing and applying image editing technologies.

F Datasheet for ULTRAEDIT

We present a Datasheet [8] for documentation and responsible usage of our internet knowledge databases. The required author statement, hosting, licensing, metadata, and maintenance plan can be found in the datasheet.

F.1 Motivation

For what purpose was the dataset created? We create this large-scale dataset to facilitate research towards image editing based on natural language instructions and regions (masks).

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? This dataset was created by Haozhe Zhao (Peking

University), Xiaojian Ma (BIGAI), Liang Chen (Peking University), Shuzheng Si (Tsinghua University), Rujie Wu (Peking University), Kaikai An (Peking University), Peiyu Yu (UCLA), Minjia Zhang (UIUC), Qing Li (BIGAI), and Baobao Chang (Peking University).

F.2 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? Yes, the dataset is publicly available on the internet.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? All datasets can be downloaded from https://huggingface.co/. Please refer to this table of URL, DOI, and licensing. The Croissant metadata can be found on the dataset hosting platform (https://huggingface.co/).

Dataset	DOI	License
ULTRAEDIT-full	10.57967/hf/2481	Creative Commons Attribution 4.0 (CC BY 4.0)
ULTRAEDIT-free-form-500k	10.57967/hf/2535	Creative Commons Attribution 4.0 (CC BY 4.0)
ULTRAEDIT-region-based-100k	10.57967/hf/2534	Creative Commons Attribution 4.0 (CC BY 4.0)

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No.

F.3 Maintenance

Who will be supporting/hosting/maintaining the dataset? The authors will be supporting, hosting, and maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? Please contact Haozhe Zhao (mimazhe55360@gmail.com), Xiaojian Ma (maxiaojian@bigai.ai) and Qing Li (liqing@bigai.ai).

Is there an erratum? No. We will make announcements if there is any.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? Yes. New updates will be posted on https://ultra-editing.github.io.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? The images in our dataset might contain human subjects, but they are all synthetic.

Will older versions of the dataset continue to be supported/hosted/maintained? Yes, old versions will be permanently accessible on huggingface.co.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Yes, please refer to https://ultra-editing.github.io.

F.4 Composition

What do the instances that comprise the dataset represent? Our data is generally stored in the Apache Parquet format, which is a table with multiple columns. We provide images (as source images and target/edited images), captions of source and target images, editing instructions, objects to be edited, metrics (CLIPimg, DINOv2, SSIM, CLIPin, CLIPout, and CLIPdir), and editing regions (optional), as separate columns.

How many instances are there in total (of each type, if appropriate)? There are ~4M samples, among which ~100K are region-based editing data, while the rests are free-form editing data.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? We provide all instances in our Huggingface data repositories.

Is there a label or target associated with each instance? No.

Is any information missing from individual instances? No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? No.

Are there recommended data splits (e.g., training, development/validation, testing)? No. The entire database is intended for training.

Are there any errors, sources of noise, or redundancies in the dataset? Please refer to Appendix E.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? The dataset is self-contained.

Does the dataset contain data that might be considered confidential? No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? We have made our best efforts to detoxify the contents via an automated procedure. Please refer to Sec. E.

F.5 Collection Process

The collection procedure, preprocessing, and cleaning are explained in detail in Section 2 of the main paper.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? All data collection, curation, and filtering are done by ULTRAEDIT coauthors.

Over what timeframe was the data collected? The data was collected between Jan. 2024 and May 2024.

F.6 Uses

Has the dataset been used for any tasks already? Yes, we have used ULTRAEDIT for training our image edit models.

What (other) tasks could the dataset be used for? Our dataset is primarily for facilitating research in building more capable image editing models that follow natural language instructions and (optionally) editing region input. Our data might also be used to benchmark existing and future image editing models.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? No.

Are there tasks for which the dataset should not be used? We strongly oppose any research that intentionally generates harmful or toxic content using our data.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, pages 8948–8957, 2019. 3

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18187–18197. IEEE, 2022. URL https://doi.org/10.1109/CVPR52688.2022.01767. 8

[3] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010. 3

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 5, 6, 8, 11, 12

[5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023. 3

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206. 13

[7] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 75–92, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72670-5. 3

[8] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. 19

[9] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems*, 19, 2006. 11, 12, 13

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *CoRR*, abs/2208.01626, 2022. URL https://doi.org/10.48550/arXiv.2208.01626. 10

[11] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. Metagpt: Meta programming for a multi-agent collaborative framework, 2023. URL https://arxiv.org/abs/2308.00352. 3

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 4

[13] Pengxiang Li, Zhi Gao, Bofei Zhang, Tao Yuan, Yuwei Wu, Mehrtash Harandi, Yunde Jia, Song-Chun Zhu, and Qing Li. Fire: A dataset for feedback integration and refinement evaluation of multimodal models, 2024. URL https://arxiv.org/abs/2407.11522. 3

[14] Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Ling-Hao Chen, Junhao Liu, Tongliang Liu, Fei Huang, and Yongbin Li. One-shot learning as instruction data prospector for large language models, 2024. URL https://arxiv.org/abs/2312.10302. 3

[15] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. URL https://doi.org/10.1007/978-3-319-10602-1_48. 3

[16] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4

[17] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aBsCjcPu_tE. 8

[18] 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. 8

[19] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR, 2022. URL https://proceedings.mlr.press/v162/nichol22a.html. 8

[20] 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, 2023. 13

[21] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 3

[22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. URL https://doi.org/10.48550/arXiv.2204.06125. 11

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

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 8, 13

[25] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[26] Christoph Schuhmann and Peter Bevan. 220k-gpt4vision-captions-from-lvis. https://huggingface.co/datasets/laion/220k-GPT4Vision-captions-from-LIVIS, 2023. 3

[27] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. *arXiv preprint arXiv:2311.10089*, 2023. 8, 11, 12, 13

[28] Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents, 2024. URL https://arxiv.org/abs/2305.13040. 3

[29] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension, 2020. 3

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=St1giarCHLP. 10

[31] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 5

[32] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2, 2014. 3

[33] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 5, 8, 11, 12

[34] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. *arXiv preprint arXiv:2303.09618*, 2023. 8

[35] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 3

[36] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023. 3