
REAEDIT: Reddit Edits As a Large-scale Empirical Dataset for Image Transformations

Supplementary Material

Table of contents

- A Data taxonomy**
 - A.1 Full taxonomy
 - A.2 Performance across edit operations
- B Data processing**
- C Additional baselines**
- D Discussion**
 - D.1 Limitations and future work
 - D.2 Social impacts
 - D.3 Ethics
- E Modeling ablations**
 - E.1. Implementation details
 - E.2. Consistency decoder
 - E.3. Data filtering
 - E.4. Processing instructions
- F. Inference time results**
 - F.1. Hyperparameters
 - F.2. Instruction rewriting
 - F.3. Quantitative evaluation on external test sets
 - F.4. Elo scores
- G Reddit experiment**
- H Edited image detection**
- I. Additional results**

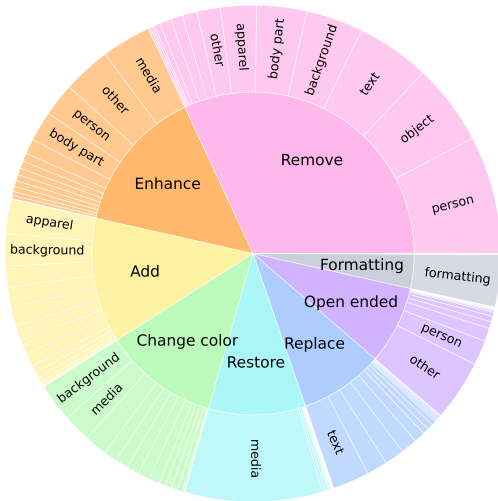


Figure 1. **Taxonomy of REAEDIT image edit requests.** There is a wide variety of task types and edit subjects, with subtle tasks like “remove” and “enhance” being the most requested.

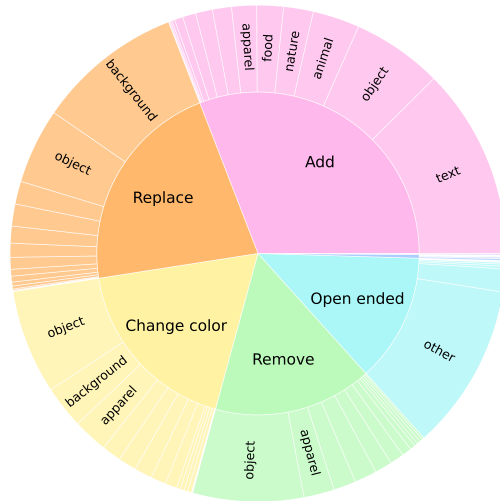


Figure 2. **Taxonomy of Emu Edit image edit requests.** There is a smaller range of task types than REALEDIT, but the distribution is fairly even.

A. Data taxonomy

A.1. Full taxonomy

We include the taxonomies of REAEDIT (Figure 1), Emu Edit (Figure 2), and MagicBrush (Figure 3) test sets, as well as the unabridged comparison between all three (Figure 4). The prompt used to taxonomize these requests is included in Figure 5. We notice REAEDIT has a more diverse set of tasks as well as a more even distribution with greater focus in tasks like “remove” and “enhance”. Emu Edit [25] has a fairly even task distribution, though a smaller set of common tasks. MagicBrush [31] has a very skewed distribution, with a high focus on “add” tasks which are not likely to be requested by human users, as humans generally include all desired elements when taking a photograph.

A.2. Performance across edit operations

We show the VIEScore comparisons of REALEDIT, AU-RORA [15], InstructPix2Pix [4] and MagicBrush [31] in Table 3. We notice that in all of the editing tasks, the REALEDIT model has the highest overall VIEScore. However, in “add” tasks, which comprise a much smaller percentage of our dataset compared to InstructPix2Pix and MagicBrush, we have a lower perceived quality, indicating that having more “add” data might improve the aesthetics. The task with the highest score for REALEDIT is “remove”, with a VIE_O score of 4.35. The “remove” task comprises the largest portion of our dataset, which may explain this

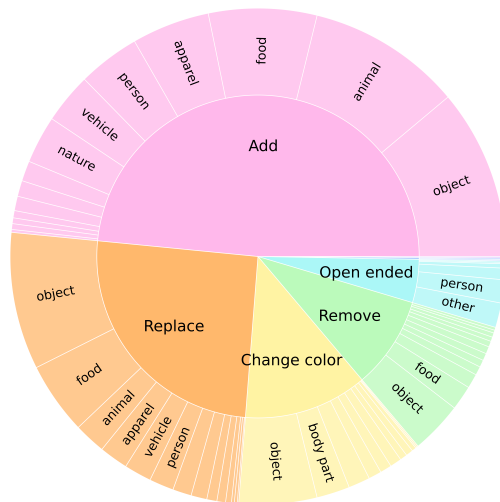


Figure 3. **Taxonomy of MagicBrush image edit requests.** There is a limited selection and extremely uneven distribution of task types, with “add” accounting for almost half of all requests.

result. The hardest task is “formatting”, the only operation for which we do not have the highest semantic completion score. This is due to the fact that this task is impossible for current models to fulfill properly, as changing file formats,

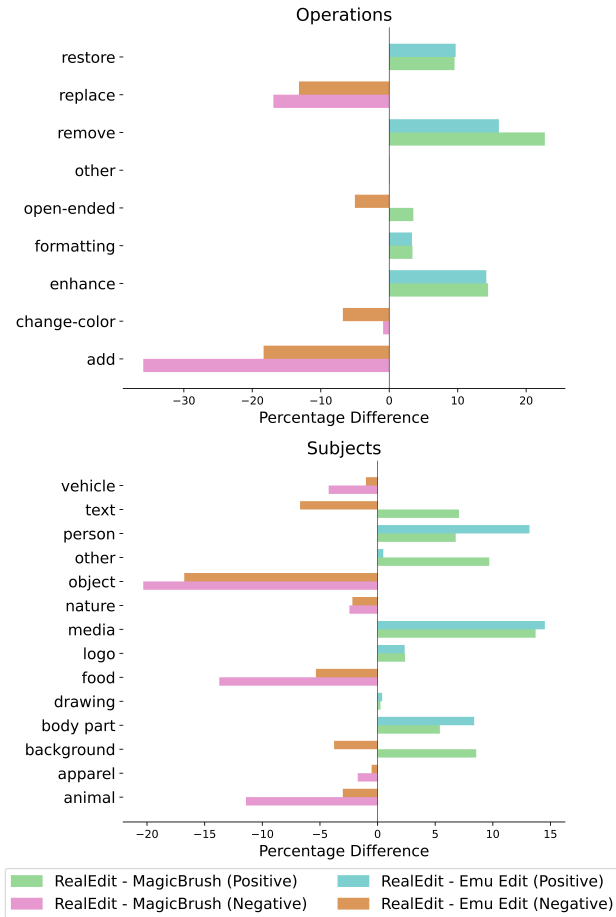


Figure 4. **Differences in the distribution** of our test set compared to MagicBrush and Emu Edit test sets. MagicBrush and Emu Edit tend to be similar in distribution to each other, but starkly different from REAEDIT.

Operation	%	Subject	%
Remove	31.9	Person	15.2
Enhance	14.5	Media	14.9
Add	12.5	Background	11.0
Change Color	11.5	Body Part	9.3
Restore	9.7	Text	8.9
Replace	8.4	Object	8.3
Open Ended	7.9	Apparel	6.5
Formatting	3.4	Format	3.4
Other	0.1	Animal	3.3
		Logo	2.5
		Vehicle	2.4
		Nature	2.2
		Other	12.1

Table 1. Distribution of edit operations in the test set.

Table 2. Distribution of edit subjects in the test set.

You are an expert at labeling image edit requests. You are great at adhering to the taxonomy provided. You are a resourceful person so you know to look at the examples for guidance.

To categorize a sample:

Step 1: select the option from the operations which best represents the task to be performed

Step 2: select the option from the subjects which best represents the subject to be edited according to the operation

Step 3: format the answer: "operation subject" If there are multiple, list each on a separate line.

Examples:

Let's assume the instruction was "Add a hat to my child."
In that case you would return Add Clothing

Let's assume the instruction was "Replace the word 'Michael Scott' on the nameplate with 'Dwight Schrute'."
In that case you would return Replace Text/Patterns

Let's assume the instruction was "Add a glowing aura to my friend."
In that case you would return Add Other

Let's assume the instruction was "humorous."
In that case you would return Open-Ended Other

You are amazing, you got this! Just remember, every request is possible to categorize according to the following taxonomy.

Here is a taxonomy of image edit requests.

Here are the possible operations:

Add: Inserting the subject into the image.

Change-Color: Color-correcting, silhouetting or otherwise changing the color of the subject, or colorizing a black and white subject.

Enhance: Sharpen, enhance, blur/unblur, remove flash/glare/lens flares.

Image-Formatting: Change file type, vectorize, adjust dimensions, etc.: any change to image parameters that do not affect the image content or aesthetics.

Open-Ended: The edit allows the editor to be creative, such as "Edit this photo." or "humorous" or "do something funny with this photo".

Remove: Erasing the subject from the image.

Replace: Substituting the subject with something specified in the instructions.

Restore: Fixing damages to the subject resulting from the preservation (e.g. stains, creases, faded color).

Other: Select this if you don't know what action is being performed.

Here are the possible subjects:

Animal: One or multiple non-human animals.

Background: The background of the image.

Body-Part: The edit is not changing an entire person, but a body part.

Clothing/Accessories: Clothing items, accessories, leashes/collars/harnesses, anything wearable by humans or animals.

Drawing: A hand-drawn drawing or handwritten note.

Food: Edible ingredients, prepared dishes, etc.

Logo: Logos or symbols.

Manmade-Structure: Buildings, furniture, other man-made structures or objects.

Media: Old photographs, screenshots, movie/game posters, memes, etc.: any form of print or digital media.

Nature: Plants, mountains, bodies of water, etc.: any naturally occurring items that are not people or animals.

Person: A person or group of people.

Text/Patterns: Text or patterns.

Vehicle: Cars, trucks, bikes, aircraft, trains, etc. any form of transportation vehicle.

Other: Select this if you don't know.

Here is an image edit request: "{{INPUT}}"

Categorize it based on the taxonomy.

Figure 5. **Prompt used for taxonomizing edit requests.** We passed this along with input images to GPT-4o.

resizing, etc. are not supported by current model architecture.

Table 3. **Breakdown of model performance by operation.** We find that our model is consistently best across all operations in VIE_O, and our strongest operation is “remove”. We use a sample of 2000 data points and take arithmetic mean of all individual scores on each data point.

Operation	AURORA			InstructPix2Pix			MagicBrush			RealEdit		
	VIE_SC	VIE_PQ	VIE_O	VIE_SC	VIE_PQ	VIE_O	VIE_SC	VIE_PQ	VIE_O	VIE_SC	VIE_PQ	VIE_O
Add	2.89	3.45	2.34	2.48	3.60	2.15	1.94	4.43	1.79	4.24	3.26	3.15
Change color	2.38	3.77	2.26	2.90	3.61	2.57	1.95	4.05	1.83	5.36	4.05	4.11
Enhance	1.86	3.00	1.88	1.80	2.91	1.79	2.41	3.44	2.33	4.73	4.03	3.95
Formatting	0.89	3.02	0.99	1.70	3.13	1.31	0.74	3.57	0.94	1.66	4.47	1.66
Open ended	2.51	2.70	1.98	2.49	3.57	2.05	2.36	3.49	1.99	4.67	2.93	3.15
Remove	2.94	4.25	2.76	1.01	3.03	1.06	2.30	4.71	2.30	5.29	5.01	4.35
Replace	2.18	3.32	1.87	2.25	3.16	1.74	1.57	3.81	1.45	3.50	3.50	2.53
Restore	1.52	2.23	1.57	1.66	2.49	1.74	1.59	2.60	1.74	4.01	2.98	3.21

B. Data processing

Test set image captioning We caption all input and ground truth images in the test set to enable evaluations with models that require captions. The process involves two main stages. First, for input image captioning, we pass the processed instruction along with the input image to LLaVA-Next[19]. This generates a caption for the input image that integrates the instruction, emphasizing key aspects of the image relevant to the editing task.

For output image captioning, we pass the input caption and edit instruction to GPT-4o, which combines these elements to generate a caption for the ground truth (edited) image, reflecting both the original content and the changes made according to the instruction. Refer to Figure 6 for examples of captions.

C. Additional baselines

After the submission of this paper, we were made aware of additional editing models with strong performance. In Table 4, we compare REALEDIT to Edit Friendly [10] and TurboEdit [6].

Table 4. **REALEDIT outperforms Edit Friendly and TurboEdit** on real-world edit requests.

Model	V _{SC} ↑	V _{PQ} ↑	V _O ↑	L1 ↓	L2 ↓	CLIP _I ↑	DINO _I ↑	CLIP _T ↑
Turbo-Edit	2.73	5.19	2.80	0.147	0.066	0.782	0.739	0.261
Edit Friendly	3.00	5.68	2.93	0.199	0.091	0.762	0.736	0.261
RealEdit	3.68	4.01	4.61	0.143	0.066	0.840	0.792	0.261

D. Discussion

D.1. Limitations and future work

REALEDIT is collected from Reddit posts from 2012-2021. As such, we have less data and a danger of it getting outdated. We plan to regularly update our dataset to ensure that the edits reflect as current culture as much as possible. This will also help in edited image detection, by facilitating the detection of edits where newer AI tools were used,

as the line between human editing and model editing is increasingly blurred.

We also filter our dataset in order to more closely match the training distribution, removing some natural diversity of human edit requests. In future work, we hope to explore different architectures capable of handling real world edit requests and editing styles.

The pretraining of the REALEDIT model uses CLIP embeddings, which while very useful for semantic changes to an image, a large portion of the REALEDIT dataset involves edits that do not involve semantic changes. Additionally, in edited image detection, some of the edits may not change the embeddings much. We urge future work to explore alternatives to such embeddings that may capture purely aesthetic changes.

D.2. Social impacts

The social impact of our dataset stems from both the effect on model training as well as the ability of our test set to be used to accurately and justly benchmark other models. The training data will inform how well the REALEDIT model performs certain types of edits. The test set on the other hand determines the factors we incentivize in other models. Accessible image editing models that are capable of handling real world tasks are extremely useful in democratizing the documentation of people’s lives. For example, some requests in REALEDIT involve restoring old photographs, many of which were paid. The REALEDIT model can help more users to document meaningful family histories, even if they cannot afford to pay for edits. We have demonstrated the efficacy of our model on making real world edits by uploading our model’s generations to Reddit. Additionally, our exploration of the contribution of REALEDIT in deepfake image detection has shown that REALEDIT increases the ability of TrueMedia.org’s ability to detect fake images, which is extremely useful in a world where images are routinely edited to cause scandals or spread misinformation.

There is a known issue in image generation models of






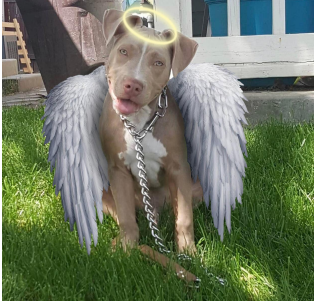


Input image	Input caption	Instruction	Output image	Output caption
	A cup of coffee on a wooden bench with a magazine and flowers.	Change the color of the cup to red with white dots.		A red cup with white dots of coffee on a wooden bench with a magazine and flowers.
	The image shows an older man and woman sitting together at a table in a restaurant.	Remove the people in the background.		The image shows an older man and woman sitting together at a table in a restaurant with no people in the background.
	A brown dog with a chain leash, sitting on the grass.	Add a halo and wings to the dog.		A brown dog with a chain leash, sitting on the grass, with a halo and wings.
	A man in a cowboy hat and bandana holding a gun, with a woman in a black dress and feather boa beside him.	Remove the lady beside the man.		A man in a cowboy hat and bandana holding a gun.

Figure 6. Examples of test set data with captions for input image and ground truth image.

generating images or making edits based on demographic biases such as smoothing wrinkles, lightening skin, and male bias in certain professions, which may offend users. Additionally, our dataset mirrors the demographic profile of Reddit users, who are predominantly Western, younger, male, and left-leaning, potentially influencing the types of images and editing requests included. We hope to study the effect of this extensively in REALEDIT in future work. There is also an issue of inappropriate edits, which we have mitigated to our knowledge in REALEDIT through filtering

of NSFW content using `opennsfw` [3], along with manual filtering in our test set.

D.3. Ethics

Some other editing datasets [31] do not use human faces in order to evade biases as well as privacy concerns. However, in REALEDIT, we determine that since over half of edit requests contain images focused on people, we must train on human data in order to be successful in completing

real world editing tasks. To mitigate privacy concerns, we use the URL in place of the actual input image so that if the original poster (OP) deletes their post, it will be removed from our dataset. We also include a form for users to request their data to be removed. This follows the standards set by RedCaps [5]. In the case of mitigating biases, we hope in future work to study the effects of using Reddit data on task completion for a wide array of demographic groups, as well as techniques or supplementary data sources to boost performance on underrepresented groups. This is a known problem in the field, and we are compelled by user preferences to include human data. Given this, although we appreciate the importance of mitigating demographic biases, this is outside the scope of a single paper.

E. Modeling ablations

E.1. Implementation details

We fine-tune the checkpoint of InstructPix2Pix [4] using the REAEDIT training set for 51 epochs on a single 80GB NVIDIA A100 GPU. The total batch size is 128, and the learning rate starts at 2×10^{-4} . We resize images to 256×256 , disable symmetrical flipping to maintain structural integrity, and apply a cosine learning rate decay to 10^{-6} over 15,000 steps with 400 warmup steps. The training process takes 24 hours.

E.2. Consistency decoder

We integrate OpenAI’s Consistency Decoder [22], which is designed to enhance the quality of specific features during inference. This has a minimal impact on overall model performance metrics but proves highly effective for improving the handling of faces, textures, and intricate patterns.

As the decoder operates independently of the underlying model, we evaluate its effectiveness with InstructPix2Pix[4] and MagicBrush[31] on a sample of 500 tasks. The results indicate that while the decoder minimally affects standard metrics, such as VIEScore[16] and CLIP-T (Table 5), it often enhances the aesthetic quality in areas requiring fine detail, such as facial reconstruction and complex textures (Figure 7).

These findings demonstrate the decoder’s potential as a lightweight, inference-only addition to improve the output quality of existing image-editing models without altering their core architectures or diffusion processes.

E.3. Data filtering

We observe that human-generated edits often introduce substantial diversity, such as rearranging objects or people, which significantly impacts Structural Similarity Index Measure (SSIM) scores. These variations create a distributional mismatch with InstructPix2Pix’s pretraining data

(Figures 8, 9, 10), where edits are generally more constrained. To better understand this difference, we analyze SSIM distributions, highlighting the gap between human edits and the structured outputs of synthetic datasets.

To make our dataset more compatible with InstructPix2Pix, we currently apply SSIM-based filtering to exclude edits that deviate too far from the pretraining distribution. Following this, we use the same CLIP-based filtering methodology employed by InstructPix2Pix to further refine the data. We verify that this filtering leads to a more capable model using the VIE-scores (Table 6) and CLIP-based metrics (Figure 11). Our approach relies on thresholding to identify and remove outliers, but we recognize that soft sampling techniques could offer a more flexible and nuanced alternative. Exploring such methods remains a promising direction for future work.

E.4. Processing instructions

Reddit users often provide vague, unclear instructions with unnecessary details, hindering the editing process. To address this, we refined these instructions for greater clarity and relevance. To evaluate the impact of this preprocessing, we trained two models under the same conditions: one with the original instructions and the other with the processed versions. Results in Table 7 and Figure 13 show that these have a significant effect on model performance.

We ran this experiment early in the development processes with a suboptimal training strategy and a smaller subset of the data, leading to much lower scores compared to our final model.

F. Inference time results

F.1. Hyperparameters

We conducted several inference-time experiments: varying the number of diffusion steps, the image and text guidance scales, and further rewriting instructions with GPT-4o to add more details.

See equation (6) in [9] for the definition of classifier-free guidance scale. The conventional wisdom is that higher image guidance scale make the generated image look more similar to the original image, while higher text guidance scale improve instruction adherence. Additionally, higher number of inference steps are believed to improve the quality of the generated image at the expense of computational time. Our statistical experiments do not capture these relationships, and even demonstrate the opposite relationship in case of image guidance scale.

Number of inference steps We observe that 20 inference steps strike a good balance between the computational time and the image quality. Specifically, we find that the average CLIP similarity between the generated image and the

Table 5. The decoder has minor effects on quantitative metrics but sometimes improves qualitative results.

Model	VIE_O	VIE_PQ	VIE_SC	L1	L2	CLIP-I	DINO-I	CLIP-T
REALEDIT w/ original decoder	3.54	3.91	4.37	0.154	0.069	0.830	0.782	0.258
REALEDIT w/ consistency decoder	3.48	3.78	4.34	0.156	0.069	0.830	0.779	0.258
Change	-0.06	-0.13	-0.03	0.002	0	0	-0.003	0
MagicBrush w/ original decoder	1.92	3.98	1.89	0.139	0.066	0.830	0.782	0.251
MagicBrush w/ consistency decoder	1.84	3.93	1.83	0.135	0.066	0.831	0.784	0.251
Change	-0.08	-0.05	-0.06	-0.004	0	0.001	0.002	0
InstructPix2Pix w/ original decoder	1.73	3.37	1.85	0.183	0.075	0.754	0.651	0.243
InstructPix2Pix w/ consistency decoder	1.89	3.40	1.95	0.180	0.073	0.758	0.648	0.244
Change	0.16	0.03	0.10	-0.003	-0.002	0.004	-0.003	0.001

Table 6. Aligning REALEDIT data to the pretraining distribution yields better results.

Model	VIE_O	VIE_PQ	VIE_SC
Filtered data	3.48	3.78	4.34
Original data	2.35	2.99	2.91

Table 7. Processing instructions improves model performance.

Model	VIE_O	VIE_PQ	VIE_SC
Processed instructions	2.42	3.72	2.84
Original instructons	2.06	3.10	2.45

most upvoted Reddit edit is approximately the same for any setting of inference steps above 20. See Figure 15 for the statistical plot and figure 14 for an example.

Text guidance scale We observe **no correlation** ($\rho = .005$) between the text guidance scale in range $[1, 14]$ and instruction adherence, as measured by CLIP similarity between the generated image and the caption describing the desired output. See Figure 17. While there is no correlation in aggregate, some individual edits may still change significantly with different text guidance scales, see Figure 16 for such an example.

Image guidance scale The generated image quality decreases sharply if the image guidance scale is above 3. Inside the $[1, 3]$ range, the image scale makes little difference in aggregate. Counter-intuitively, we observe a **negative** correlation ($\rho = -.106$) between image guidance scale and CLIP similarity between the input and generated images. In other words, higher image guidance values result in **less similar** images on average, which contradicts conventional assumptions about guidance scales and warrants further in-

vestigation. See Figure 18.

F.2. Instruction rewriting

As the diffusion model lacks reasoning capabilities, it often fails when asked to interpret abstract or creative instructions. To improve outcomes on these examples, we employ a large language model (LLM) to rewrite instructions in a more specific manner, similar to Dalle-3 [2]. Since only creative edit tasks benefit from this technique, we do not make this part of our main pipeline. We gave the input image and the original instruction to GPT-4o with the prompt “You are given an image editing instruction. If the instruction is already concrete and specific, do not rewrite it at all. If the instruction is vague or does not make sense for the image, then rewrite it. Make the new instruction specific and detailed, e.g. do not use words ‘enhance’, ‘adjust’, ‘any’.”

F.3. Quantitative evaluation on external test sets

Despite being out of distribution, the REALEDIT model performs comparably to other models on the synthetic datasets Emu Edit [25] and MagicBrush [31]. On several metrics (VQA_CLIP and TIFA on MagicBrush and VQA_Llava, VQA_Flan-t5 and TIFA on Emu Edit), the REALEDIT model is within 1 standard deviation of the highest scoring model, indicating that it is fairly generalizable to new tasks.

F.4. Elo scores

To evaluate Elo scores, we leverage Amazon Mechanical Turk (MTurk) for conducting pairwise comparisons. We selected 200 diverse examples from our dataset to ensure coverage of various editing tasks and performed comparisons across all seven models in our benchmark. This process resulted in a total of 4,200 pairwise evaluations, providing a robust dataset for assessing human preferences. We present a table of pairwise winrates (Figure 20)

In addition to evaluating our dataset, we extended our analysis to the Imagen Hub Museum[17] tasks, building on

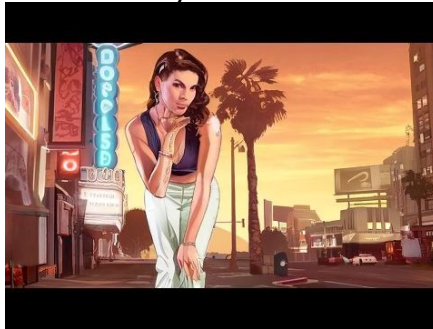
Remove text and logo



Original decoder



Consistency decoder



Lighten the face of the guy on the left



Original decoder



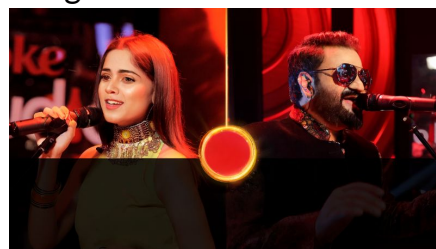
Consistency decoder



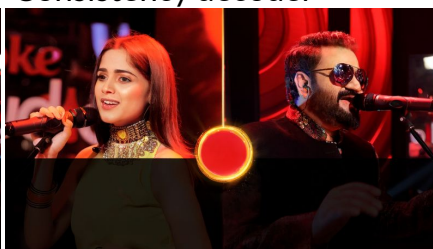
Remove the text



Original decoder



Consistency decoder



Make the sky blue



Original decoder



Consistency decoder



Figure 7. Consistency decoder allows for more aesthetic generation of faces.

the results from the GenAI Arena[11]. Using their generations, available on HuggingFace, we incorporated results from our model to facilitate direct comparisons. For these evaluations, we conducted a new round of pairwise comparisons where we matched one model from their benchmark against our model for the same tasks. This allowed us to

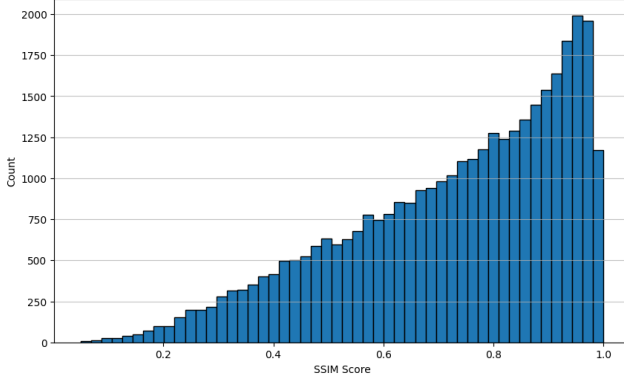


Figure 8. SSIM distribution of InstructPix2Pix training data.

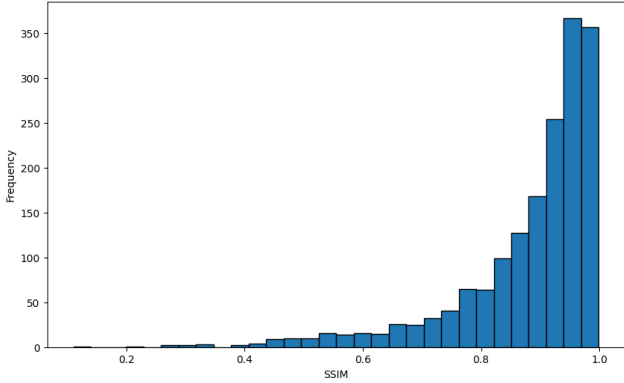


Figure 9. SSIM distribution of MagicBrush training data.

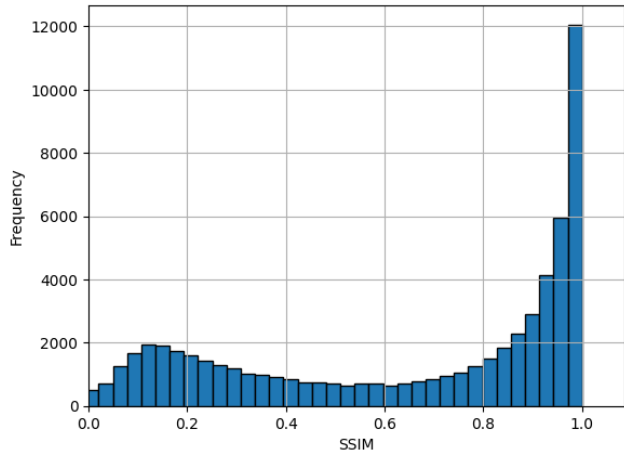


Figure 10. SSIM distribution of REALEDIT training data.

directly assess how our model performs relative to state-of-the-art models on external datasets.

The evaluations on MTurk followed a structured protocol to ensure reliability and consistency. Workers were asked to compare image outputs based on task completion, realism, and alignment with instructions. The use of MTurk enabled us to gather diverse human feedback efficiently and at scale. The full results are presented in Table 9, highlighting the comparative performance across different models.

G. Reddit experiment

To evaluate the generalization capability of our model, we deployed it on Reddit. Specifically, we targeted two subreddits: r/PhotoshopRequest and r/estoration, which focus on image editing and restoration tasks. Adhering to the community guidelines of these subreddits, we collected posts requesting image edits and processed them using our model.

For each processed request, we submitted a comment containing the generated output image along with a brief message asking for feedback from the original poster (OP). With this experiment, we gathered qualitative evaluations from humans, and provide insight into the model’s performance in real world scenarios. See Figures 22 and 23.

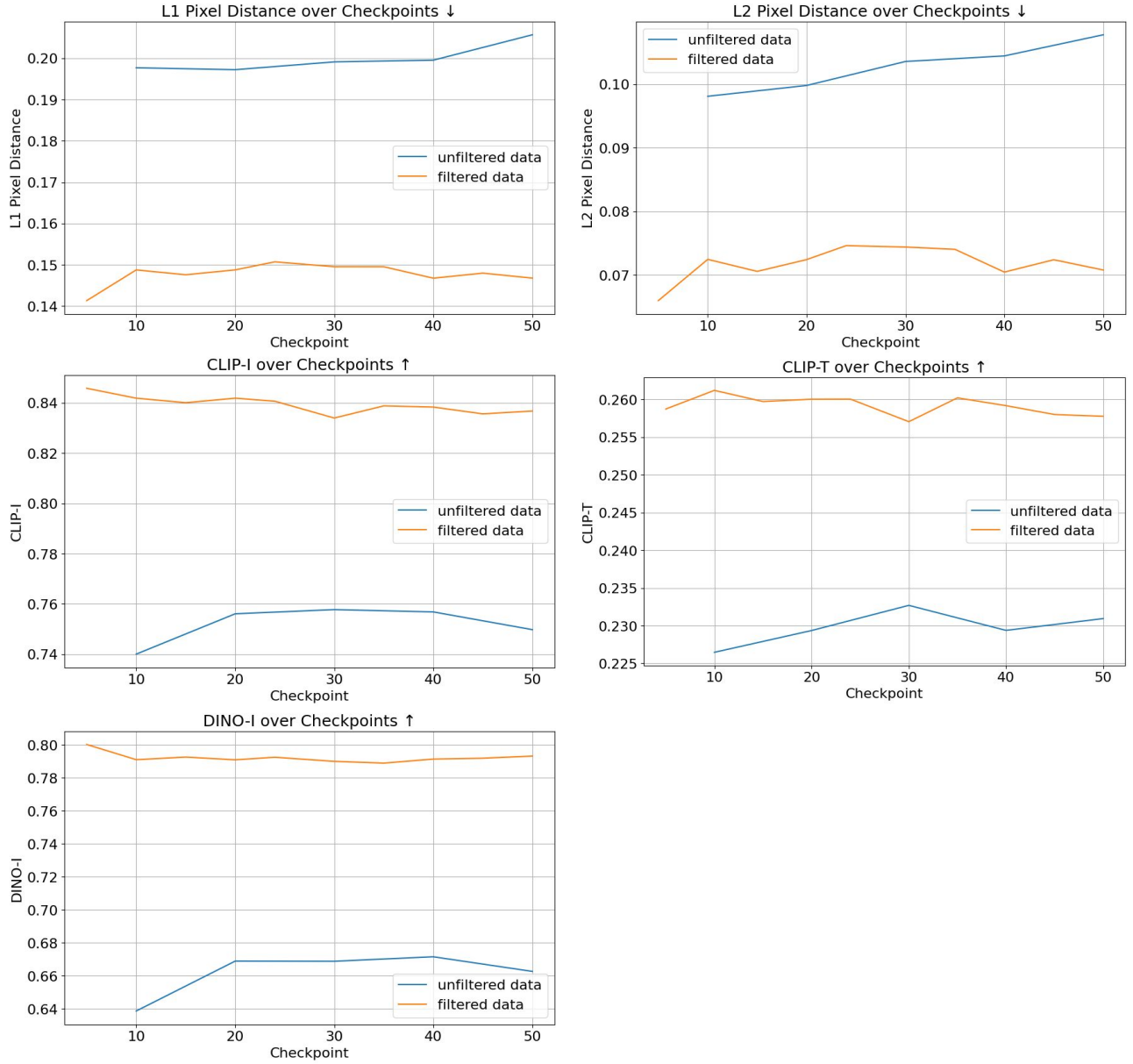


Figure 11. Filtering the data massively improved CLIP-based metrics.

Table 8. **Evaluation on MagicBrush and Emu Edit test sets.** All scores within 1 standard deviation of the highest score are underlined. The REALEDIT model is still able to perform competitively on some metrics despite these tasks being out of distribution.

Model	MagicBrush Test Set						Emu Edit Test Set					
	VIES.SC ↑	VIEPQ ↑	VIEO ↑	VQA.llava ↑	VQA.CLIP ↑	TIFA ↑	VIES.SC ↑	VIE.PQ ↑	VIE.O ↑	VQA.llava ↑	VQA_Flan-t5 ↑	TIFA ↑
AURORA [15]	4.11	3.86	5.52	0.5179	<u>0.6517</u>	<u>0.6968</u>	3.40	<u>4.86</u>	3.81	<u>0.4923</u>	<u>0.6178</u>	0.6705
Emu Edit [25]	N/A	N/A	N/A	N/A	N/A	N/A	4.66	<u>5.11</u>	5.72	0.5130	0.6489	<u>0.6692</u>
HIVE [32]	2.86	5.02	3.43	<u>0.5200</u>	<u>0.6547</u>	<u>0.6918</u>	1.89	5.50	2.06	0.4372	0.5258	0.6447
InstructPix2Pix [4]	2.63	<u>4.70</u>	3.06	0.4490	0.5518	0.6615	2.15	<u>5.00</u>	2.36	0.4261	0.5061	0.6343
MagicBrush [31]	<u>3.43</u>	<u>4.89</u>	4.11	0.5554	0.7138	0.7103	2.91	<u>5.47</u>	3.13	0.4680	0.5808	<u>0.6628</u>
Null-text Inv. [21]	2.77	<u>4.74</u>	3.29	<u>0.5246</u>	<u>0.6429</u>	<u>0.6899</u>	3.43	<u>5.10</u>	3.93	0.4823	0.5931	<u>0.6578</u>
SDEdit [20]	0.90	2.26	1.02	0.4185	0.4191	0.6167	0.95	3.23	1.06	0.4406	0.5145	0.6417
RealEdit	3.12	3.60	4.09	0.5088	<u>0.6299</u>	<u>0.6865</u>	3.27	<u>4.86</u>	3.84	<u>0.4938</u>	<u>0.6158</u>	<u>0.6650</u>

Simplify the given image editing instruction. Remove URLs, typos, irrelevant details, and expressions of gratitude. Summarize the main task and be concise. Your output should be a concise image editing request. If you think the request is humorous or ambiguous, classify it as 'humorous'.

Examples of good input and outputs:

Input instruction: [Specific] Can someone remove the text? I wanna use it as a mobile wallpaper. (J5)
Output instruction: Remove the text.

Input instruction: My friend's mom has a birthday coming up, and hoping to get her childhood photo restored.
Output instruction: Restore this photo.

Input instruction: [SPECIFIC] I've been asked for a headshot-- can you make this look like one? (please!)
Output instruction: Turn this image into a professional headshot.

Input instruction: Please photoshop me in anyway you want. I just want it to be funny.
Output instruction: humorous.

Input instruction: {{INPUT}}
Output instruction:

Figure 12. GPT-4o prompt for instruction rewriting.

Table 9. Elo scores of models based on the GenAI[11] test set.

Model	Elo Rating	95% C.I.	Sample Size
MagicBrush [31]	1107	-39/+47	132
CosXLEdit [1]	1064	-49/+42	133
RealEdit	1043	-12/+17	1117
InfEdit [30]	1023	-44/+39	122
InstructPix2Pix [4]	1011	-50/+47	117
Prompt2prompt [8]	1011	-46/+46	119
PNP [27]	992	-43/+62	122
SDEdit [20]	991	-48/+35	126
CycleDiffusion [29]	933	-41/+49	120
Pix2PixZero [23]	834	-46/+41	126

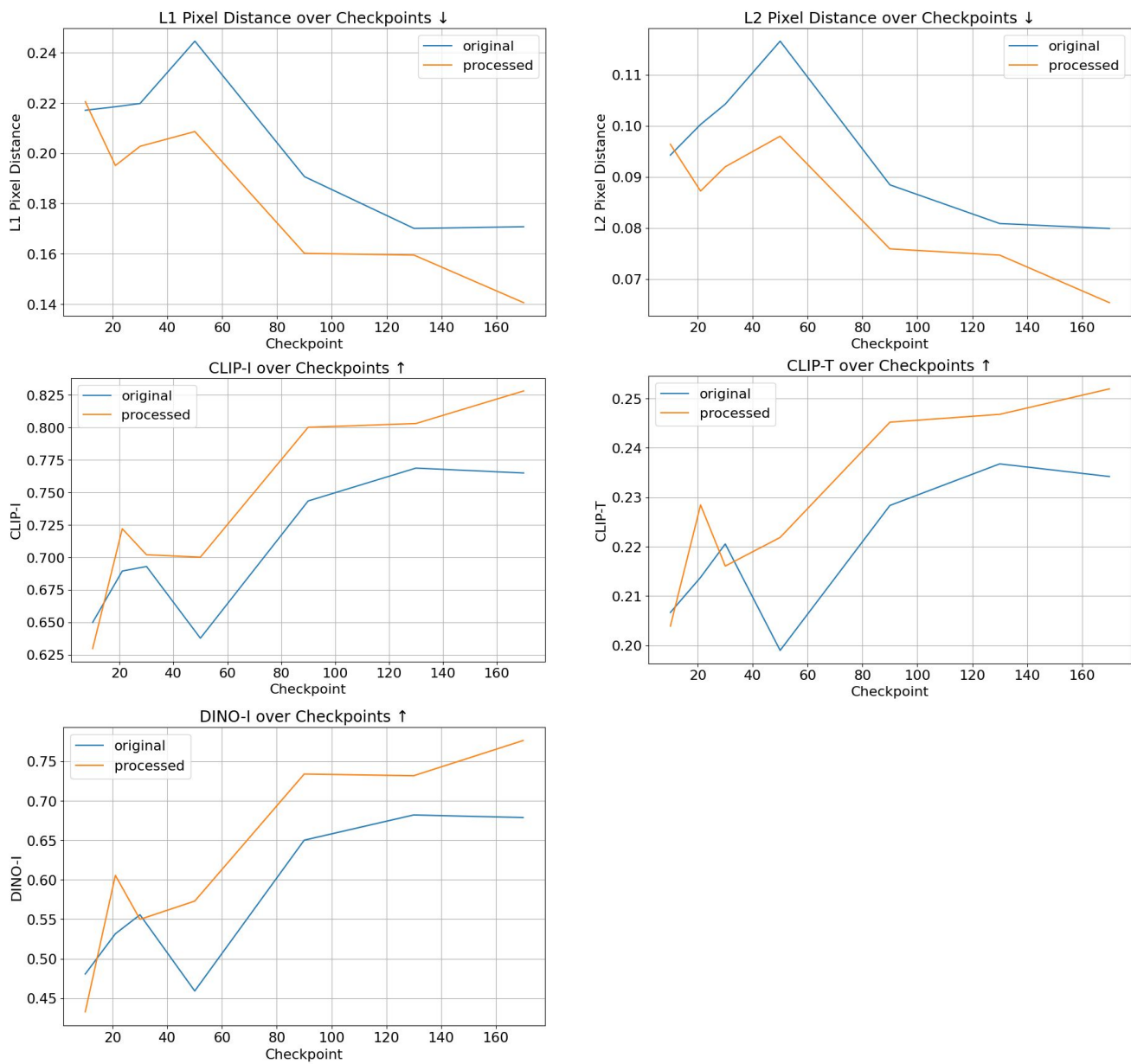


Figure 13. Processing instructions consistently yields better results on CLIP-based results.

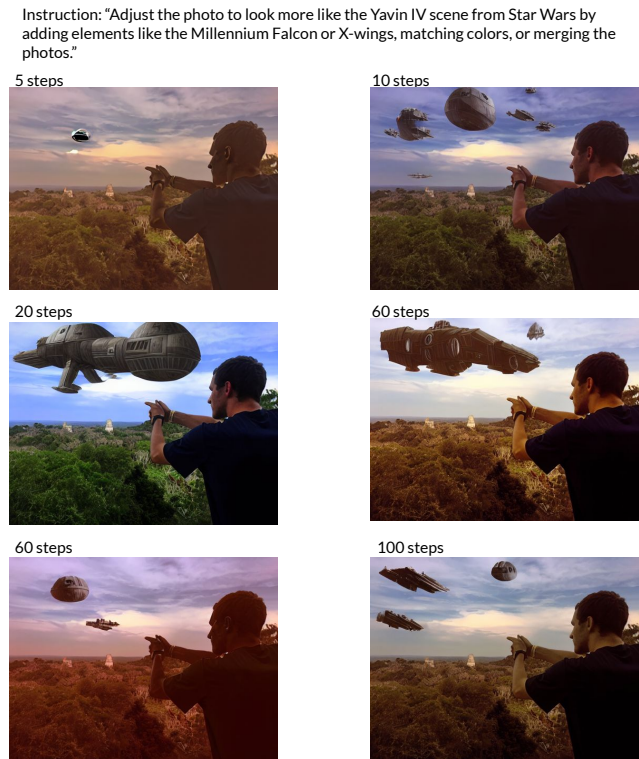


Figure 14. Increasing the number of diffusion steps above 20 usually does not improve the quality.



Figure 16. An example where guidance scales behave as expected.

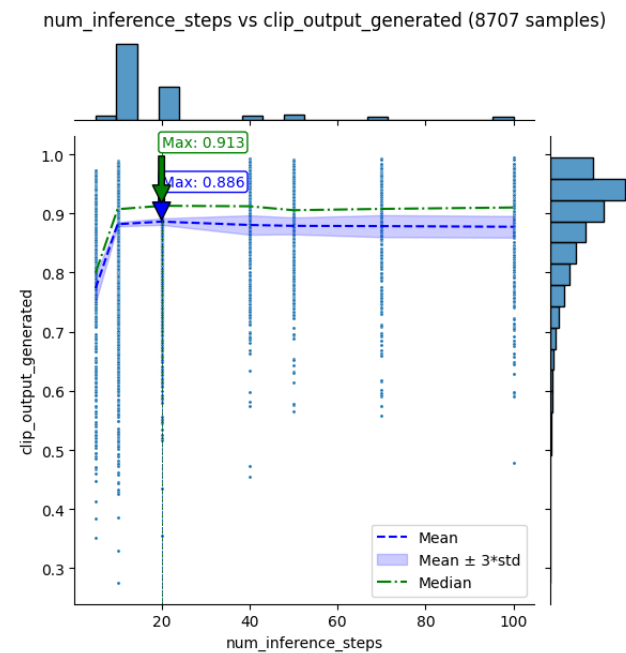


Figure 15. The number of inference steps does not improve the generated image quality, as measured by the CLIP similarity between the generated image and the most upvoted Reddit edit.

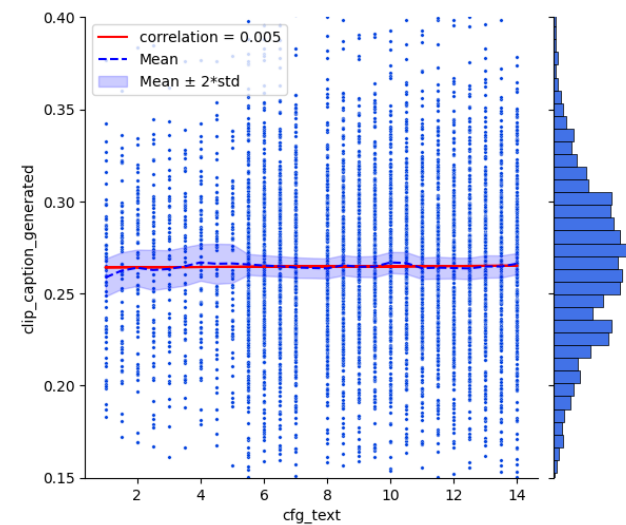


Figure 17. Text guidance scale has no effect on instruction adherence, as measured by the CLIP similarity between the generated image and the caption of the expected output, as in figure 6.

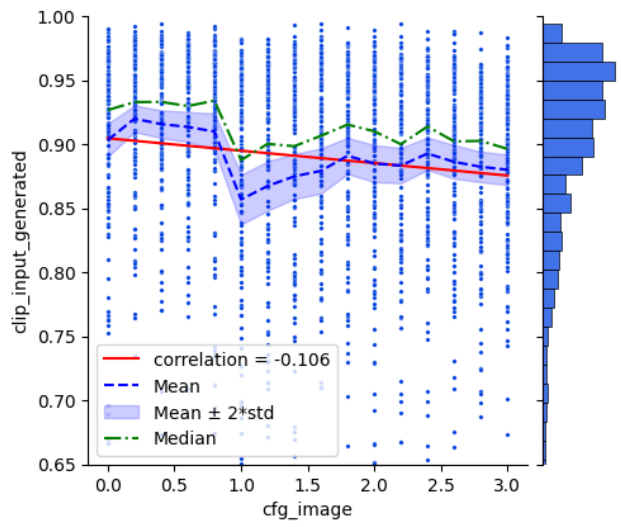
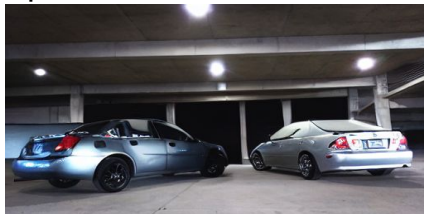


Figure 18. Increased image guidance scale results in **less** similar images, as measured by CLIP similarity between the input and generated images.

Original image



Put the car and people in space.



Place the cars in a space background, maintaining their position. Add stars in the background.



Original image



Remove the person photobombing in the background.



Remove the reflection of the photographer visible in the goggles.



Original image



Flip the colors of this guitar.



Swap the black and orange colors on the guitar body.



Figure 19. Detailed instructions can improve edit quality on certain classes of tasks.

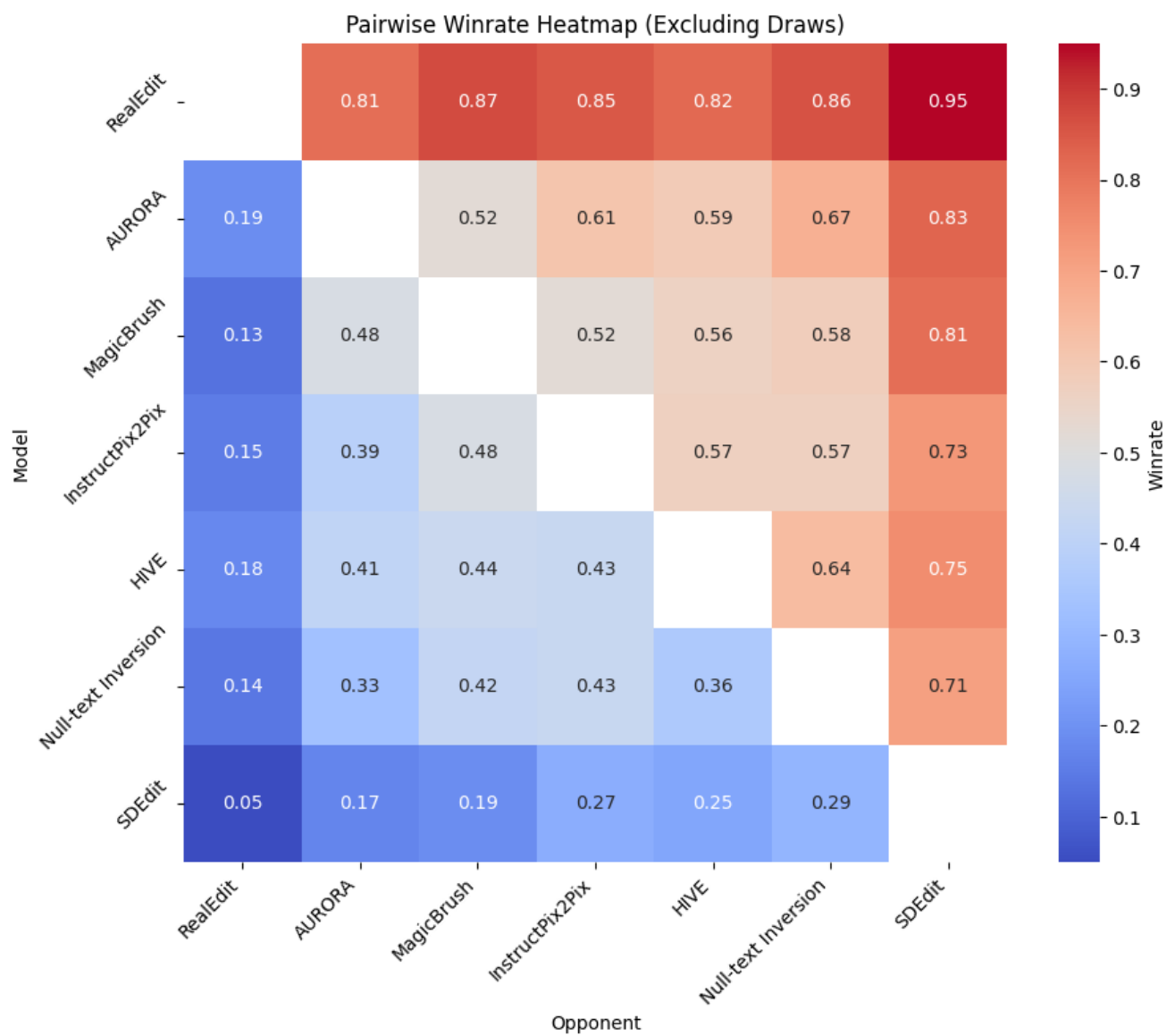



Figure 20. **Heatmap of pairwise winrates on our test set.** We excluded draws for this heatmap.


Input Image




Edit Instruction

Remove the power lines.

Left Edit



Right Edit



Please evaluate the images and select exactly one option below.

☐ Left image is better

☐ Right image is better

☐ Both images are equally good

☐ Both images are equally bad

Figure 21. **Interface for Elo evaluation on MTurk.** To complete Elo evaluations, we hired workers on Amazon Mechanical Turk to compare the quality of different editing models.

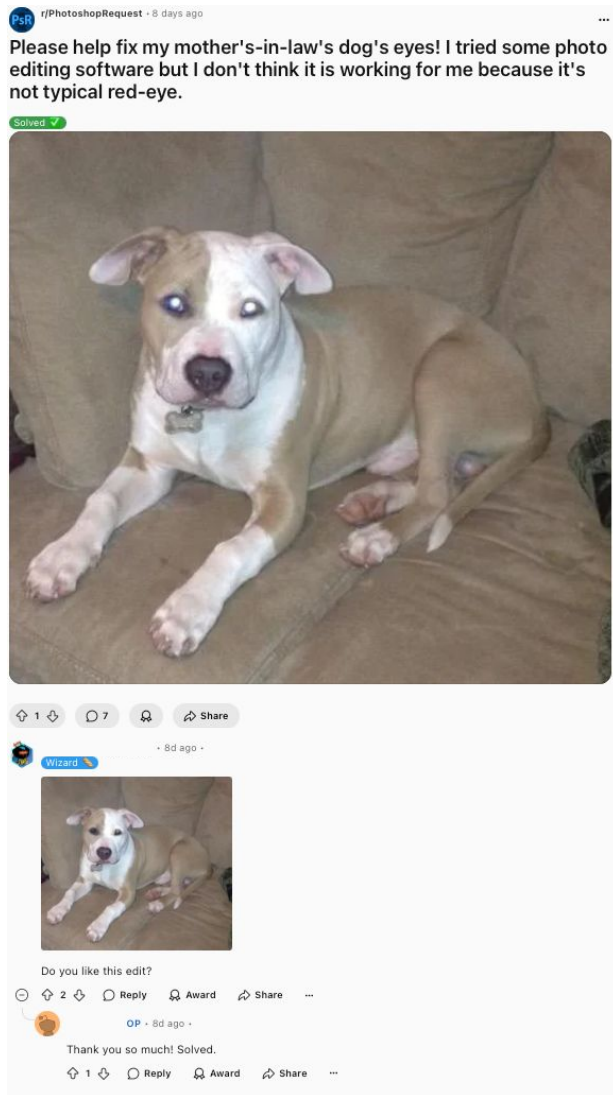


Figure 22. **Our model successfully completes new requests on Reddit.** Deployed on the original subreddits, it handled in-the-wild requests effectively as seen by OP's response.

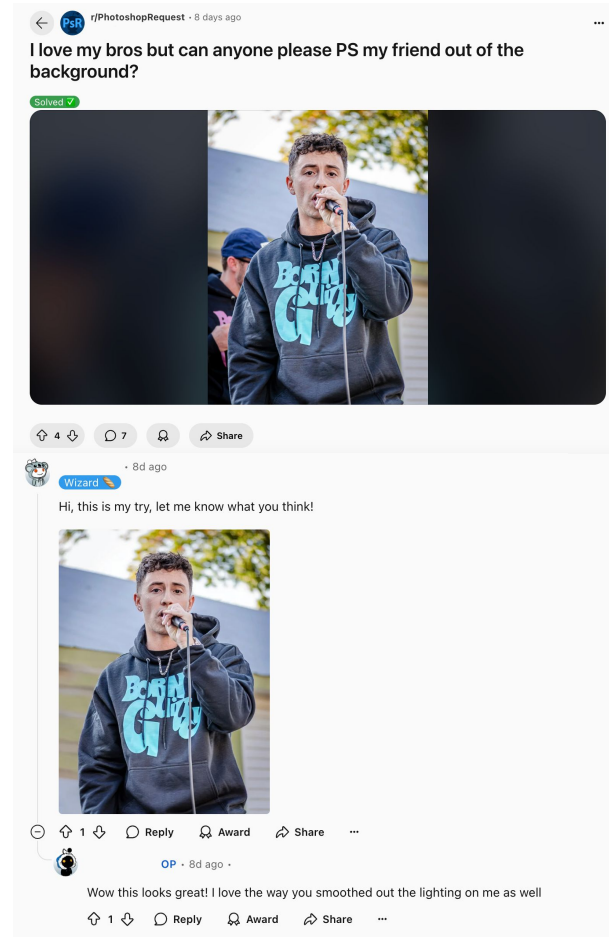


Figure 23. **Our model successfully completes new requests on Reddit.** Deployed on the original subreddits, it handled in-the-wild requests effectively as seen by OP's response.

H. Edited image detection

Data processing and training The baseline classifier undergoes a multi-stage training process: initially on academic datasets and subsequently fine-tuned on TrueMedia.org’s proprietary data. In total, the baseline model is trained on 65K images with a near equal 50/50 split between real and generated images. To assess the value of REALEDIT data for fake image detection, we train a second version of UFD by combining the original data with REALEDIT data. Specifically, we include only photographs, excluding non-photographic images such as digital artworks, screenshots, cartoons, and infographics, filtered using GPT-4o. This single-stage training incorporates an additional 37K original and 37K edited images, resulting in a total of 139K images.

In the first stage of training, the TrueMedia.org model took over 24 hours to train on an A10G GPU with 20GB of RAM and the remaining three stages took 4 hours. Our optimized model took 1.5 hours to train on a L40S GPU with 40GB of RAM.

Table 10. Breakdown of fake image sources in the training recipe of the TrueMedia.org model used as our baseline.

Source	Count
DiffusionDB [28]	16K
StyleGAN2-FFHQ [14]	8K
Stable-Diffusion-Face [26] (512 resolution)	2.4K
Stable-Diffusion-Face (768 resolution)	2.4K
Stable-Diffusion-Face (1024 resolution)	2.4K
Fakes uploaded to TrueMedia.org	2K

Table 11. Breakdown of real image sources in the training recipe of the baseline model.

Source	Count
CelebA-HQ (Reals) [12]	23K
Random sample of COCO-Train-2017 [18]	5K
Flickr-Faces-HQ Dataset (FFHQ) [13]	3K
Reals uploaded to TrueMedia.org	0.7K

TrueMedia.org’s in-the-wild test set TrueMedia.org’s in-the-wild test set includes images uploaded between 8/16/2024 and 11/10/2024. We randomly sample 100 real images and then sample 100 fake images selected from those tagged as “likely photoshopped” by professional sleuths in TrueMedia.org’s media database, ensuring the



Figure 24. Top: An edited image that inserted a bear to make it seem the camera crew was being chased. Bottom: Grad-CAM heat-map visualization highlighting the regions of attention.

evaluation focuses on human-edited images rather than exclusively AI-generated edits. Tool usage data was available for some images, revealing that approximately 80% of the fake images were human edits created with Photoshop, while the remaining 20% involved human edits combined with AI tools such as Dream Studio AI, Insightface AI, and Remaker AI.

Qualitative example To understand how the classifier operates, we use Gradient-weighted Class Activation Mapping (Grad-CAM) [24] to analyze an example. In Figure 24, we show an edited image where a bear was added to the background using Photoshop. The original image did not include the bear. The baseline model incorrectly classified this photo as unedited, whereas the classifier trained with REALEDIT data correctly identified it as edited. Grad-CAM highlights the areas of the image most influential to the classifier’s decision, as seen in the figure, where the focus is on the region around the bear. The specific implementation we adapted is from Gildenblat and contributors [7].

I. Additional results

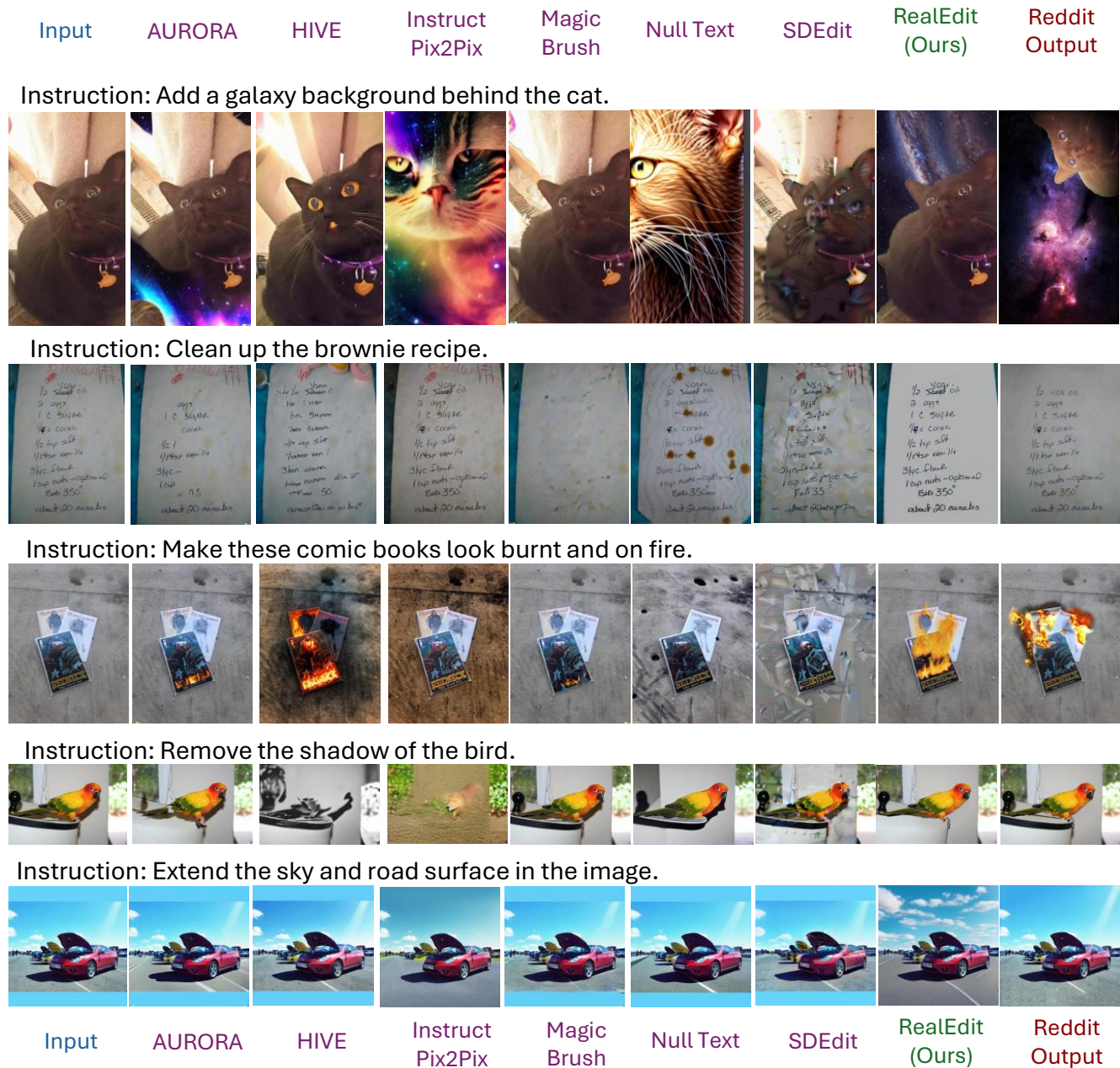


Figure 25. **Additional examples of REALEDIT generations on REALEDIT test set** compared to all other baseline models. We notice that the REALEDIT model consistently outperforms other models in task completion as well as aesthetic quality.

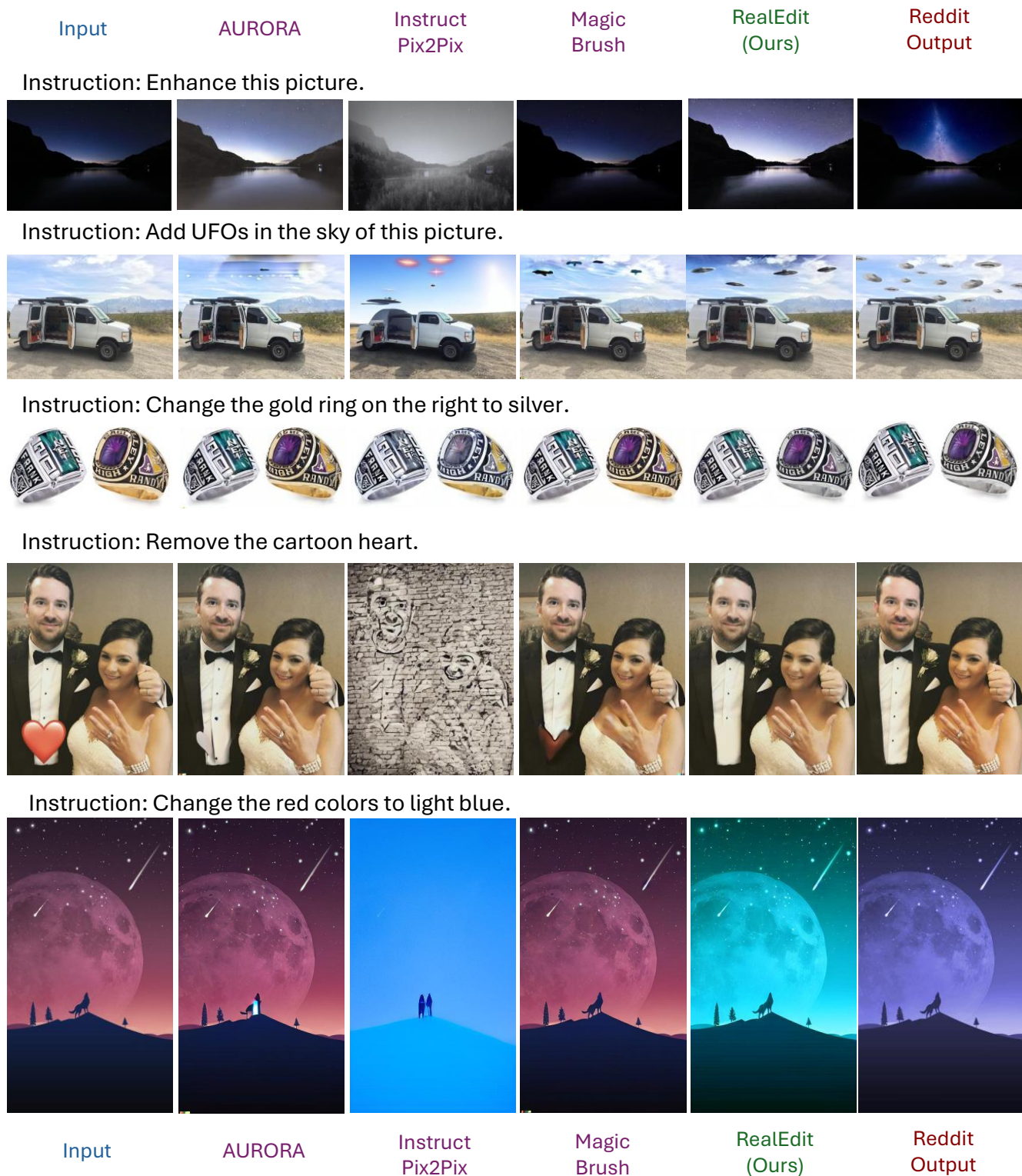


Figure 26. **Additional examples of REALEDIT generations on REALEDIT test set** compared to select high performing baseline models. We notice that the REALEDIT model consistently outperforms other models in task completion as well as aesthetic quality.

References

- [1] Stability AI. Cosxl. <https://huggingface.co/stabilityai/cosxl>, 2024. Accessed: 2024-11-05.
- [2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023.
- [3] bhky. opennsfw2. <https://github.com/bhky/opennsfw2>, 2020. Accessed: 2024-04-27.
- [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] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021.
- [6] Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. Turboedit: Text-based image editing using few-step diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024.
- [7] Jacob Gildenblat and contributors. Pytorch library for cam methods. <https://github.com/jacobgil/pytorch-grad-cam>, 2021.
- [8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.
- [9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [10] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12469–12478, 2024.
- [11] Daya Jiang, Muchen Ku, Tong Li, Yajie Ni, Shu Sun, Rui Fan, and Wei Chen. Genai arena: An open evaluation platform for generative models. *arXiv preprint arXiv:2406.04485*, 2024.
- [12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018.
- [13] 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.
- [14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020.
- [15] Benno Krojer, Dheeraj Vattikonda, Luis Lara, Varun Jampani, Eva Portelance, Christopher Pal, and Siva Reddy. Learning action and reasoning-centric image editing from videos and simulations. *arXiv preprint arXiv:2407.03471*, 2024.
- [16] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhui Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation, 2023.
- [17] Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhui Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. In *The Twelfth International Conference on Learning Representations*, 2024.
- [18] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.
- [19] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.
- [20] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jianjun Wu, Jun-Yan Zhu, and Stefano Ermon. Sedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.
- [21] 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.
- [22] OpenAI. ConsistencyDecoder. <https://github.com/openai/consistencydecoder>, 2023. Accessed: April 27, 2024.
- [23] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.
- [24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2019.
- [25] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024.
- [26] tobecwb. Stable diffusion face dataset. <https://github.com/tobecwb/stable-diffusion-face-dataset>, 2023. Accessed: 2024-04-02.
- [27] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.
- [28] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models, 2023.
- [29] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7378–7387, 2023.

- [30] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language. 2024.
- [31] Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024.
- [32] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9026–9036, 2024.