

UniBench: Fully Automated Evaluation of Image Editing Models

Suho Ryu Dongmin Choi Jongook Yoon
Graduate School of Data Science, Seoul National University, Korea
{jmhera2007, chrisandjj, jonguki}@snu.ac.kr

Abstract

Recently, numerous text-guided image editing methods have emerged, utilizing the remarkable capabilities of large-scale diffusion-based generative models. However, a reliable and efficient evaluation protocol for comparing these methods is lacking. To address this problem, we introduce UniBench, a standardized and fully automated benchmark for quantitative evaluation of text-guided image editing models. UniBench features a curated dataset of images, annotations with editable instances, candidate edit targets, and edit prompt templates for each image. It also includes an automated evaluation pipeline that employs pre-trained instance segmentation models, vision-language models, and perceptual similarity measures to score the fidelity and consistency of generated images from each edit type. Using UniBench, we benchmark two cutting-edge diffusion-based editing methods: IMAGIC and Prompt-to-Prompt. Our results show that UniBench’s automated evaluation pipeline correlates with the edit quality of the results and effectively identifies the shortcomings of editing models. We believe our benchmark may clear the way to developing reliable text-guided image editing tools in the future.

1. Introduction

Triggered by recent development of powerful and highly utilizable diffusion based text-to-image generation models such as Stable-Diffusion [12], IMAGEN [13] and DALL-E [11], research on their applications in various fields is rapidly expanding over the last few years. One prominent application of these highly acclaimed diffusion models is in text conditioned image editing technology. Various diffusion-based image editing models [3, 5, 7] surged into research field in the early years, but recently their development has reached a plateau.

One of the major reasons of stagnant research pool is lack of fine benchmarks to accurately and objectively evaluate these image editing models. Developing such good benchmark for image editing models is quite challenging because there is no definitive answer for image edit-

ing tasks, and the direction of editing can vary in countless ways depending on the original image. Developing such good benchmark for image editing models is extremely difficult due to ambiguity in non-existent answer in image editing tasks, and boundlessness and volatility of possible approach of editing depending on the original image.

There have been a few notable attempts in this field. Shi et al. [15] proposed a benchmark baseline for image editing task which provided a set of editing requests with corresponding original and ground truth result image for a model to process and evaluate itself. However, editing requests were limited to those which clear ground truth was able to be given such as removing an object in an image or grey scaling an image. Kawar et al. [5] and Wang et al. [16] also presented their own unique benchmark guideline and image set, but its scoring method relied on human survey. Basu et al. [1] diversified and structured edit categories (types) and target object classes for equal assessment throughout wide range of edit operations, but it was still in the limit of human evaluation.

Here we present ‘UniBench’, a novel benchmark pipeline and dataset for image editing models, with diverse range of editing operation types and target objects which completely objective and automatic in edit task generation and model result evaluation. In section 3, we provide detailed explanation of the dataset, edit caption annotation and evaluation model. In section 4 and 5 we give simple test experiments of UniBench on some of State-of-the-Art editing models in the field.

2. Related Works

Text-Guided Image Editing Models. Recently, text-guided image diffusion models have shown exceptional image generation capabilities, achieving state-of-the-art FID [4] scores on benchmarks like MS-COCO. Typically pre-trained on extensive datasets of image-text pairs such as LAION [14] using a diffusion objective, these models have advanced significantly. Beyond generating images, they are now being utilized to edit real images, marking a significant leap in their application. In this paper we evaluate some of State-of-the-Art text-guided image editing diffusion mod-

els: IMAGIC and Prompt-to-Prompt.

Image Editing Benchmarks. So far, benchmarks such as TedBench [5], EditBench Wang et al. [16], and EditVal Basu et al. [1] have been introduced for text-guided image editing, but each has its limitations. Their scoring methods depend on human surveys. EditBench specifically evaluates only on mask-guided image editing methods, necessitating an additional mask along with the edit prompt. In contrast, our proposed UniBench can be applied to any text-guided editing method and is fully automated in both edit task generation and model evaluation.

3. UniBench

UniBench is composed of three components: editable image dataset, edit caption annotations and evaluation modules. The image dataset provides set of images well-suited to image editing tasks along with lists of instance segmentation annotations of the images. Edit caption annotations are dictionary of all possible candidate edit information and prompts paired with each image in the dataset. Evaluation module evaluates and scores the output of a model to be evaluated.

3.1. The Dataset

Base Dataset. Images and annotation pairs are carefully selected from MS-COCO [6] dataset. MS-COCO provides image captions, pre-defined instance classes and instance segmentation annotations which are key components of edit command and prompt generation. Pre-annotated instance list of each image allows easy selection of object to be edited. Diverse pre-defined instance classes provide excellent list of object classes to use as edit target objects. Image captions are crucial to many image editing models, and they are also reformed as target prompt templates in later work. Presence of many well-performing instance segmentation models trained on MS-COCO set allows convenient construction of evaluation module.

Data Selection. We manually select total 50 images with 1. diverse, unbiased clearly distinguishable instances that can be edited, 2. captions that include all the edit candidate instances and 3. high degree of freedom for any editing tasks. Images with various class instances, each unique in kind, are preferred for diversity and minimal confusion when locating the instance to edit(or that has been edited). There must exist a caption that describes the instance to edit since majority of image manipulation models are conditioned with text prompt tweaked from a corresponding image caption. Images with higher edit-ability, like a photo of a book on a field, is more suitable than those with limited edit-ability, like a photo of a book in a book shelve, since it provides greater diversity in edit task selection.

Data Pre-processing. We crop and resize selected images and annotations into required input size of image editing

models. Operators manually review each image, locate instances within each image, square-crop and resize the images to desired size without harming any distinguish-ability of an instance. We also process any annotations that is required to be removed, translated, resized or cropped together with the image.

3.2. Edit Prompt Generation

Edit Task Definition. We define four types of instance editing tasks that can be applied to general images and instances. *Change class* task selects an instance from an original image and replace it with different class object. *Change color* task changes color of an instance without harming its semantics and shapes. *Move* task shifts the position of an instance within the original image to a location defined relative to another object instance (hereafter called an anchor). *Generate* task creates an instance on a location described based on another anchor. Examples of each task type are shown in figure 1.

Edit Caption Annotation. Figuring out which instance in an image is editable to what kind of target class, color or position is ambiguous and can vary from image to image. Thus, for each image, we first manually select and list editable instances. Then, for each instance, we annotate with information needed for each edit task. For *change class* task, we list instance classes that are physically reasonable to replace the original object. For *change color* task, we list possible colors to change to. For *move* task, we label a list of objects that can be used as an anchor object together with list of plausible relative positions from each anchors. Lastly, for *generate* task, a list of instance classes to generate, list of anchors with relative positions are labeled. We exclude class and color identical to original instance, and we also avoid any classes that are already present in the image for clear location of edited instance during the evaluation. We also provide caption templates for each "original instance"- "edit type" pair, where a model can insert any class, color or position from annotated list to generate an edit caption/prompt.

LLM Utilization and Revision. In the process of generating an edit prompt, manually creating lists of candidate target and prompts is inefficient and time-consuming. To resolve the problem, we utilize Large Language Model [GPT-4o 2, 9]. We provide example image and annotation pairs as a guideline for the GPT-4o to follow. Given a set of images and names of instances in the images, we request GPT-4o to generate desired annotations. If generated annotations are inappropriate, we provide additional prompting to ensure proper result is generated. GPT-4o occasionally generates candidate edit targets that do not belong to given object class library. Thus, after the generation we go through final human revision before it is used in model evaluation. We list some examples of edit caption annotation in figure 1.


Image	Original object	Edit type	Candidate target	Candidate anchor / position	Prompt templates
	"cat"	Change class	["dog", "bird", "teddy bear", ...]	-	["A {} sitting on a suitcase, ...]
		Change color	["black", "brown", "white", ...]	-	["A {} cat sitting on a suitcase, ...]
		move	-	["suitcase"] / ["in front of", "next to", "behind", ...]	["A cat sitting {position} a {anchor}", ...]
	"suitcase"	Change class	["backpack", "couch", "bed", ...]	-	["A cat sitting on a {}", ...]
		Change color	["green", "brown", "black", ...]	-	["A cat sitting on a {} suitcase", ...]
		generate	["banana", "bird", "book", ...]	["suitcase"] / ["above", "next to", ...]	["A cat sitting on a suitcase, and a {target} {position} a {anchor}", ...]

Figure 1. Edit caption annotation example.

3.3. Evaluation Pipeline

We generate edit prompts with annotated edit captions, and we process original dataset images through image manipulation models with generated prompts to produce edited output images. To evaluate the testing models, we score the model outputs into two categories: fidelity and consistency.

Figure 2 shows detailed process of editing output evaluation for the results of each task type. We process output images through pre-trained instance segmentation network to locate edit target object and anchor objects. If any of the object is not found within the image, locations (segmentation mask and bounding box) of original objects from original image is used instead except for cases where target object location cannot be predicted from original object location (*e.g.* locating target object from *move* task). We then separate input and output images into original/target/anchor objects and the backgrounds.

Fidelity score consists of two sub-scores: target fidelity and position fidelity. Target fidelity score is defined as text-to-image semantic alignment between target class and separated target object with CLIP [10] embedding similarity. Position fidelity score is computed only for *move* and *generate* tasks and is defined as CLIP similarity between position phrase (*e.g.* '{target}' in front of {anchor}') and output image cropped to only contain target and anchor object.

Consistency score also consists of two sub-scores: target consistency and background consistency. Target consistency is computed only for *change color* and *move* tasks where it is important to preserve the semantics and shapes of an object. For *change color* results, target consistency is calculated through $(1 - \text{LPIPS} [17] \text{ perceptual similarity})$ between grey scaled image of separated original and target object. For *move* results, it is obtained from CLIP similarity between original and target object. Meanwhile, background consistency is computed as $(1 - \text{LPIPS distance})$ between original and edited background images.

We average all scores among the same kind within same edit type group to obtain mean scores for each edit type. Then We average along task types to get total fidelity and

consistency score of the model. Every evaluation scores range from 0 (bad) to 1 (good), and due to adoption of average reduction, all reduced scores also range from 0 (bad) to 1 (good).

Usage of instance segmentation model allow us to automatically locate and separate edited object with provides key foundation of benchmark automation. By specifying detailed edit types and scoring fields, UniBench may provide detailed and precise evaluation of a model. Through applying CLIP alignment scores only between separated single object image and short text description, we expect it to overcome the limitation of VLM, failing to interpret complex relations between multiple objects. LPIPS perceptual similarity is expected to provide comprehensive and semantic measure of consistency of image elements that should not be changed.

4. Experiments

IMAGIC [5] is one of SOTA diffusion based text prompt conditioned image editing model which takes input set of original image, target prompt and edit strength η . Here we compare the UniBench scores of IMAGIC output with different edit strength parameter. Prompt-to-prompt [3, hereafter P2P] proposed a novel image editing framework utilizing cross-attention control enabling editing of diffusion synthesized image. By the help of null-text inversion technique [8], the scope of P2P was further expanded not only for synthesized images but also for any given real images. Here we employ P2P with null-text inversion to edit UniBench image sets and evaluate output results.

Throughout all 50 images of the dataset, for every editable instances in an image, one edit task/prompt per task type were generated with randomly selected target class, color, anchor object, position and caption template from annotated sets. As a result, around 4 - 10 edit tasks per image (depending on the number of instances and possible edits) and 238 tasks in total were generated. Edit results of each task were drawn with both IMAGIC with 4 different η values (0.8, 0.9, 1.0, 1.1) and P2P with null-text inversion.

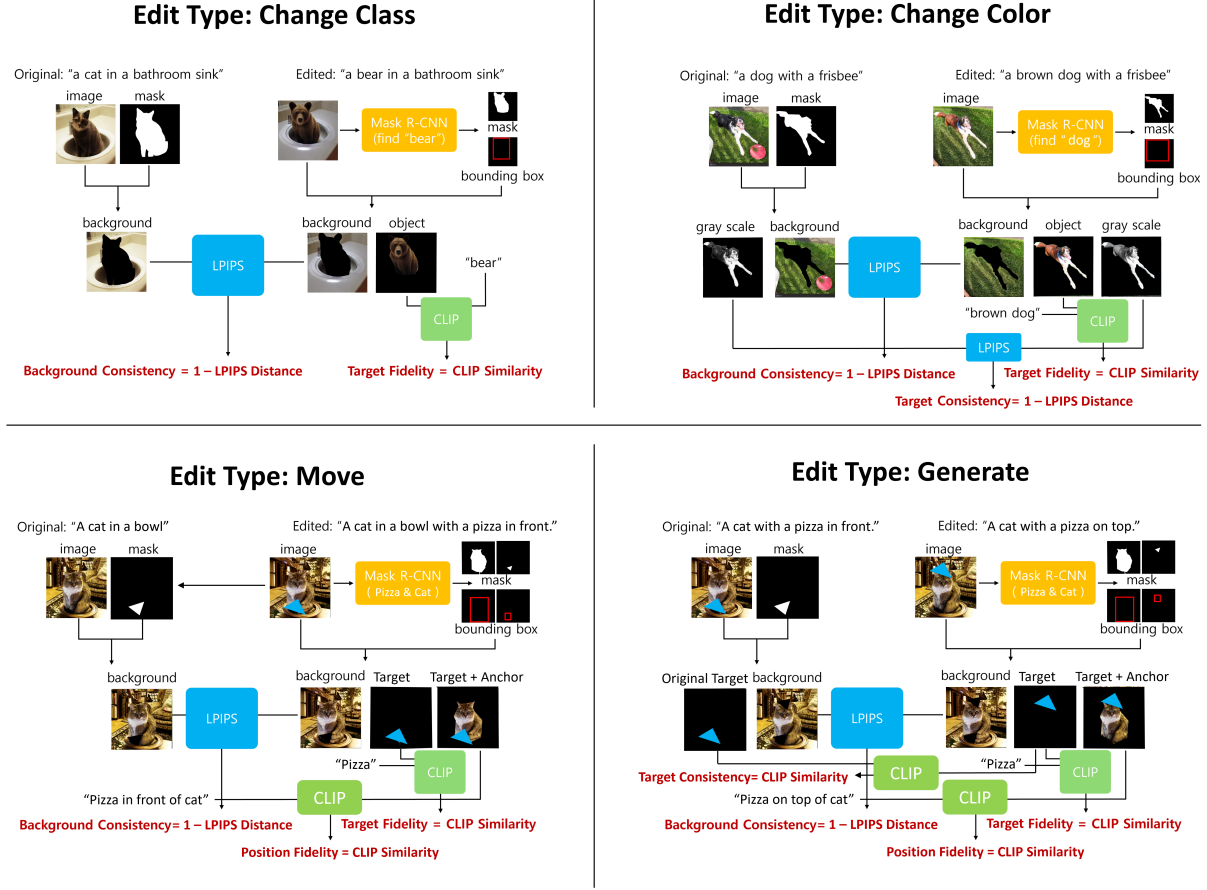


Figure 2. Evaluation process for each task.

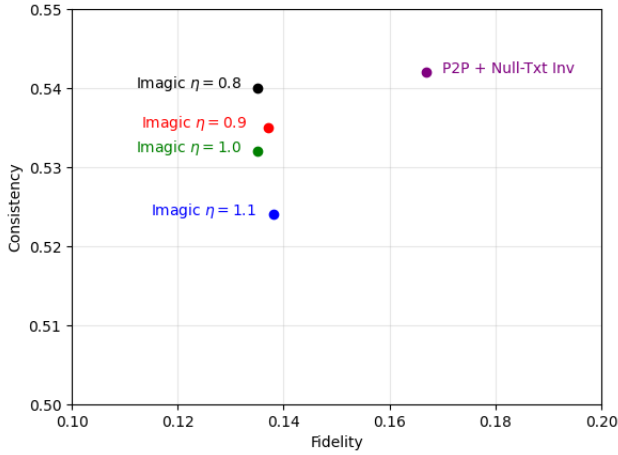


Figure 3. UniBench results of IMAGIC outputs with four different edit strength ($\eta = 0.8, 0.9, 1.0, 1.1$) and P2P with null-text inversion.

5. Results

UniBench results for all the models are listed in table 1 and plotted in figure 3. Overall, fidelity scores of all the models reside far below 0.3 implying failure of majority of edit tasks. It can also be noticed in figure 4 that many results fail to accomplish the task. Position fidelity scores tend to be higher than the others. This may due to the anchor object since regardless of the edit result, anchor object is likely to present in both image and prompt raising the CLIP similarity score. IMAGIC has failed to generate detectable target object in *generate* task showing zero fidelity score which is given when detection model cannot find the target object (also shown in figure 4). P2P with null-text inversion also shows near zero result for generate task. Within IMAGIC models, $\eta = 1.1$ case shows highest fidelity score which sounds reasonable recalling that η is edit strength parameter. However, deviations between models are too small to tell statistical significance.

Consistency scores tend to be higher relative to fidelity scores. The result well agrees with edit examples shown in 4

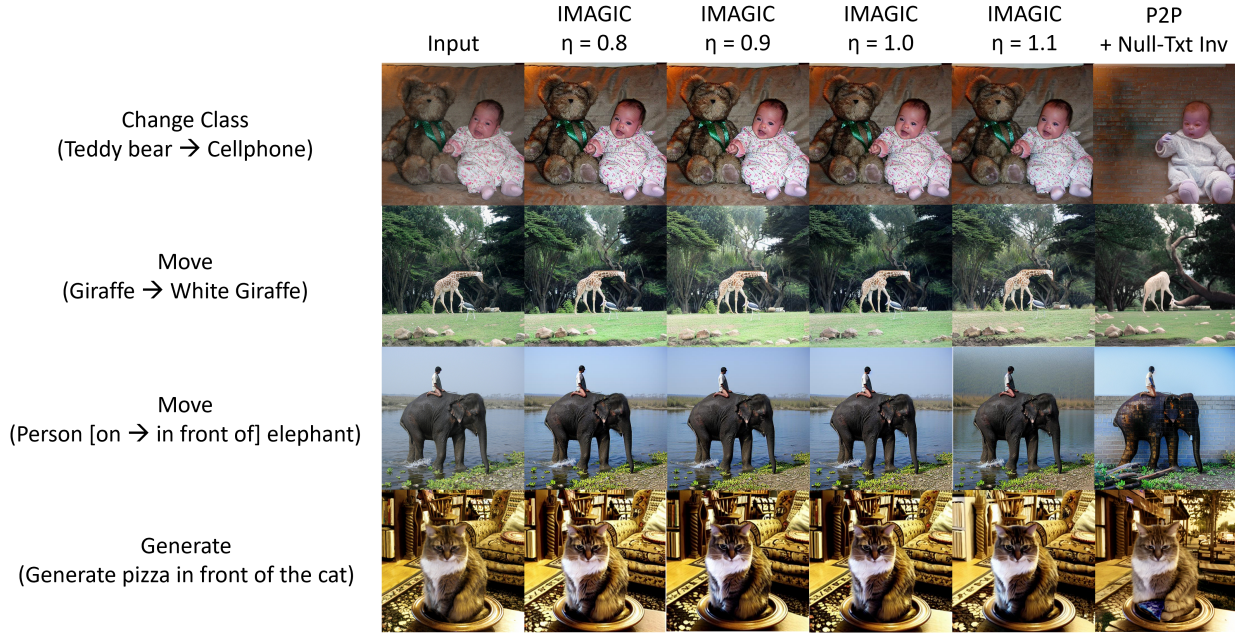


Figure 4. One random sampled example input/output image pair per task type generated by IMAGIC with four different edit strength ($\eta = 0.8, 0.9, 1.0, 1.1$) and P2P with null-text inversion.

where we can see great consistencies between original and edited image. Within IMAGIC models, consistency score decreases as η increases which can be predicted from both meaning of η and results in 4. The potential of Unibench as elaborate evaluation framework can be found from its successful performance distinguishment of IMAGIC models.

Except for *move* task scores, all evaluated scores of P2P exceed those of IMAGIC models. This discernment of UniBench may enable objective comparison between many text-conditioned image editing frameworks.

6. Conclusion & Discussion

An image-annotation pair set containing set of images suitable for editing tasks along with their instance segmentation annotations were carefully sampled from MS-COCO dataset. For each image, possible editing tasks with corresponding text prompt templates were generated through LLM and were revised by human annotators. With created image-annotation set we were able to successfully generate sample edit tasks and guide prompts. Created tasks were carried out with image editing models that are to be evaluated (IMAGIC with several edit strength and Prompt-to-prompt with Null-text inversion). Finally, model results were scored with our UniBench evaluation pipeline without any difficulties. Quantitative measures by Unibench revealed yet insufficient performance of image editing models and were able to give delicate comparison between different models.

From the experiment results in section 5, we can seek the potential of UniBench as a novel benchmark of image manipulation models which can quantitatively evaluate a model in various specific categories and task types together with overall performance without any human evaluation. With reliable evaluation metric for image editing frameworks, which has been absent until today, truly meaningful development of image manipulation models will finally be possible.

Due to lack of computation resources and time, we were only able to run limited number of editing operations. The reliability of the evaluation need to be improved with greater number of edit results per model. Furthermore, number of independent trials need to be taken on the average and variations of the results need to be taken to provide stable result along with its errors which is crucial for statistical analysis. We still lack in diversity of edit task types, and we leave it as a future objective. We adopted simple average reduction of evaluated scores; however, careful and logical scaling and weighing between evaluation criteria needs to be done. Finally, the result set size is still absurdly too small to implement any generation quality assessment like FID. Overall, it is necessary to enlarge the scale of Unibench pipeline.

References

- [1] Samyadeep Basu, Mehrdad Saberi, Shweta Bhardwaj, et al. EditVal: Benchmarking Diffusion Based Text-Guided Image Editing Methods. *arXiv e-prints arXiv:2310.02426*, 2023. 1,

		Fidelity			Consistency		
		target fidelity	position fidelity	average	target consistency	background consistency	average
IMAGIC $\eta = 0.8$	change class	0.102	-	0.102	-	0.921	0.921
	change color	0.144	-	0.144	0.744	0.932	0.838
	generate	0.000	0.000	0.000	-	0.928	0.928
	move	0.385	0.746	0.566	0.947	0.944	0.946
	total			0.203			0.908
IMAGIC $\eta = 0.9$	change class	0.105	-	0.105	-	0.907	0.907
	change color	0.148	-	0.148	0.736	0.929	0.833
	generate	0.000	0.000	0.000	-	0.918	0.918
	move	0.380	0.757	0.569	0.944	0.937	0.941
	total			0.205			0.900
IMAGIC $\eta = 1.0$	change class	0.109	-	0.109	-	0.902	0.902
	change color	0.141	-	0.141	0.744	0.918	0.831
	generate	0.000	0.000	0.000	-	0.923	0.923
	move	0.358	0.767	0.563	0.917	0.931	0.924
	total			0.203			0.895
IMAGIC $\eta = 1.1$	change class	0.124	-	0.124	-	0.882	0.882
	change color	0.157	-	0.157	0.731	0.911	0.821
	generate	0.000	0.000	0.000	-	0.905	0.905
	move	0.363	0.728	0.546	0.914	0.908	0.911
	total			0.207			0.880
P2P + Null-Txt Inv	change class	0.235	-	0.235	-	0.952	0.952
	change color	0.192	-	0.192	0.789	0.948	0.869
	generate	0.051	0.072	0.062	-	0.923	0.923
	move	0.339	0.692	0.516	0.844	0.930	0.887
	total			0.251			0.908

Table 1. UniBench results of IMAGIC outputs with four different edit strength ($\eta = 0.8, 0.9, 1.0, 1.1$) and P2P with null-text inversion.

- 2
- [2] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language Models are Few-Shot Learners. *arXiv e-prints arXiv:2005.14165*, 2020. 2
- [3] Amir Hertz, Ron Mokady, Jay Tenenbaum, et al. Prompt-to-prompt image editing with cross attention control. *arXiv e-prints arXiv:2208.01626*, 2022. 1, 3
- [4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *arXiv e-prints arXiv:1706.08500*, 2017. 1
- [5] Bahjat Kavar, Shiran Zada, Oran Lang, et al. Imagic: Text-Based Real Image Editing with Diffusion Models. *arXiv e-prints arXiv:2210.09276*, 2022. 1, 2, 3
- [6] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft COCO: Common Objects in Context. *arXiv e-prints arXiv:1405.0312*, 2014. 2
- [7] Chenlin Meng, Yutong He, Yang Song, et al. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv e-prints arXiv:2108.01073*, 2022. 1
- [8] Ron Mokady, Amir Hertz, Kfir Aberman, et al. Null-text Inversion for Editing Real Images using Guided Diffusion Models. *arXiv e-prints arXiv:2211.09794*, 2022. 3
- [9] OpenAI, Josh Achiam, Steven Adler, et al. GPT-4 Technical Report. *arXiv e-prints arXiv:2303.08774*, 2023. 2
- [10] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning Transferable Visual Models From Natural Language Supervision. *arXiv e-prints arXiv:2103.00020*, 2021. 3
- [11] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, et al. Zero-Shot Text-to-Image Generation. *arXiv e-prints arXiv:2102.12092*, 2021. 1
- [12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, et al. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv e-prints arXiv:2112.10752*, 2021. 1
- [13] Chitwan Saharia, William Chan, Saurabh Saxena, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv e-prints arXiv:2205.11487*, 2022. 1
- [14] Christoph Schuhmann, Richard Vencu, Romain Beaumont, et al. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *arXiv e-prints arXiv:2111.02114*, 2021. 1
- [15] Jing Shi, Ning Xu, Trung Bui, et al. A benchmark and baseline for language-driven image editing. *arXiv e-prints arXiv:2010.02330*, 2020. 1
- [16] Su Wang, Chitwan Saharia, Ceslee Montgomery, et al. Imagen Editor and EditBench: Advancing and Evaluating Text-Guided Image Inpainting. *arXiv e-prints arXiv:2212.06909*, 2022. 1, 2
- [17] Richard Zhang, Phillip Isola, Alexei A. Efros, et al. The

Unreasonable Effectiveness of Deep Features as a Perceptual
Metric. *arXiv e-prints arXiv:1801.03924*, 2018. 3