I^{2}EBench: A Comprehensive Benchmark for Instruction-based Image Editing

Yiwei Ma[1*] **Jiayi Ji**[1*] **Ke Ye**[1] **Weihuang Lin**[1] **Zhibin Wang**[2†]
Yonghan Zheng[1] **Qiang Zhou**[2] **Xiaoshuai Sun**[1†] **Rongrong Ji**[1]
[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, 361005, P.R. China.
[2] Inf Tech Company, Hangzhou, 310000, P.R. China.
yiweima@stu.xmu.edu.cn xssun@xmu.edu.cn

Abstract

Significant progress has been made in the field of Instruction-based Image Editing (IIE). However, evaluating these models poses a significant challenge. A crucial requirement in this field is the establishment of a comprehensive evaluation benchmark for accurately assessing editing results and providing valuable insights for its further development. In response to this need, we propose **I^{2}EBench**, a comprehensive benchmark designed to automatically evaluate the quality of edited images produced by IIE models from multiple dimensions. I^{2}EBench consists of 2,000+ images for editing, along with 4,000+ corresponding original and diverse instructions. It offers three distinctive characteristics: 1) *Comprehensive Evaluation Dimensions:* I^{2}EBench comprises 16 evaluation dimensions that cover both high-level and low-level aspects, providing a comprehensive assessment of each IIE model. 2) *Human Perception Alignment:* To ensure the alignment of our benchmark with human perception, we conducted an extensive user study for each evaluation dimension. 3) *Valuable Research Insights:* By analyzing the advantages and disadvantages of existing IIE models across the 16 dimensions, we offer valuable research insights to guide future development in the field. We will open-source I^{2}EBench, including all instructions, input images, human annotations, edited images from all evaluated methods, and a simple script for evaluating the results from new IIE models. The code, dataset and generated images from all IIE models are provided in github: https://github.com/cocoshe/I2EBench.

1 Introduction

Instruction-based Image Editing (IIE)Brooks et al. [2023], Geng et al. [2023], Zhang et al. [2024a], Li et al. [2023c], Wang et al. [2023b], Zhang et al. [2023a], Fu et al. [2024], which aims to edit an image using a text instruction, provides a user-friendly way for the community to edit images. Over the past few years, significant progress has been made in IIE, with the development of diffusion models Ho et al. [2020], Sohl-Dickstein et al. [2015], Welling and Teh [2011], Kulikov et al. [2023] and large vision-language models (LVLMs) Liu et al. [2023a,b], Fei et al. [2024c,a,b], Ma et al. [2024]. However, there is a pressing need for a comprehensive benchmark to effectively assess the performance of these models. An ideal evaluation framework should not only measure the editing quality across different dimensions but also align with human perception to ensure reliable measurements. Furthermore, the evaluation should highlight the specific strengths and weaknesses of each model, thereby offering valuable insights for future endeavors in data selection, training

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

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

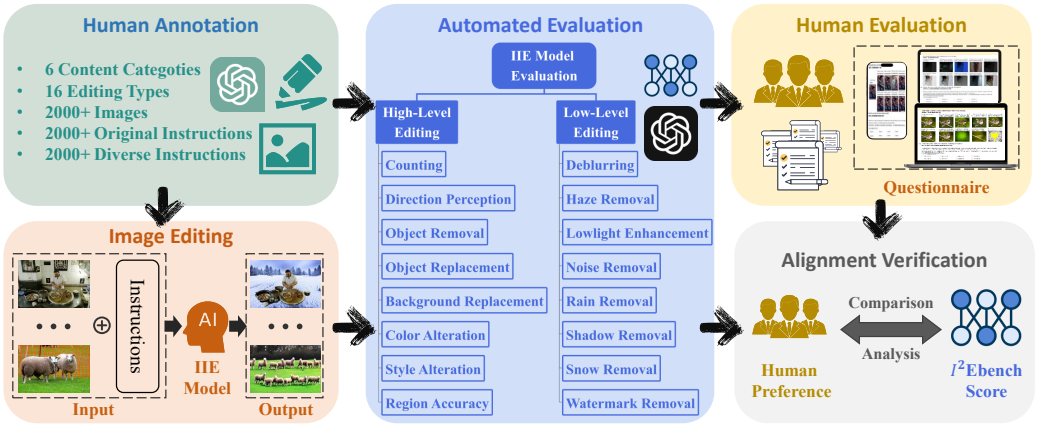


Figure 1: Overview of I^{2}EBench, an automated system for evaluating the quality of editing results generated by instruction-based image editing (IIE) models. We collected a dataset of over 2000+ images from public datasets Lin et al. [2014], Guo et al. [2023b], Martin et al. [2001], Chen et al. [2021], Ancuti et al. [2019], Liu et al. [2021b,a], Qu et al. [2017], Nah et al. [2017], Shen et al. [2019], Wei et al. [2018] and annotated them with corresponding original editing instructions. To diversify the instructions, we used ChatGPT Achiam et al. [2023] to generate varied versions. With the collected images and the original/diverse editing instructions, we utilized existing IIE models to generate edited images. Subsequently, we developed an evaluation methodology to automatically assess the adherence of edited images to the provided instructions under different dimensions. We also implemented human evaluation to obtain human preferences for editing results of different IIE models. Finally, we analyzed the correlation between automated evaluation and human evaluation, confirming alignment with human perception.

strategy selection, and architecture design within this field. However, evaluating an IIE model poses challenges due to the diverse range of editing types and the inherent difficulty in assessing the level of alignment between edited images and given instructions.

Existing evaluation metrics for IIE could be divided into three categories: 1) conventional metric; 2) user study; 3) benchmark. The first category Brooks et al. [2023], Geng et al. [2023], Zhang et al. [2024a], Li et al. [2023c], Wang et al. [2023b], Huang et al. [2024b] employs conventional metrics to evaluate IIE models, including CLIP Score Radford et al. [2021], CLIP Text-Image Direction Similarity Radford et al. [2021], PSNR Korhonen and You [2012], SSIM Wang et al. [2004], and LPIPS Zhang et al. [2018]. The advantage of this approach is its ease of use. However, a single metric is not suitable for evaluating all types of editing. For instance, CLIP score measures the similarity between images and text, making it less suitable for low-level visual editing tasks like denoising and low-light enhancement. Similarly, PSNR, which measures image similarity, is not adequate for high-level visual editing tasks such as object removal and replacement. The second category Li et al. [2023c], Zhang et al. [2023a], Fu et al. [2024] involves methods that evaluate the effectiveness of different techniques by soliciting ratings from human participants. This approach directly reflects human preferences and aligns the results with human perception. However, it is a costly method and lacks reproducibility, as the test sets and participants may be not consistent in each evaluation. The final category comprises benchmarks Kawar et al. [2023], Wang et al. [2023c], Basu et al. [2023], Huang et al. [2024a] specifically designed for evaluating IIE models. While these benchmarks are tailored for IIE, they have certain limitations. For example, TedBench Kawar et al. [2023] evaluates only 100 images with commonly occurring editing types, which may not sufficiently demonstrate the capabilities of IIE models. EditBench Wang et al. [2023c] focuses on mask-guided editing, rendering it unsuitable for evaluating mask-free methods. In EditVal Basu et al. [2023], only a limited set of dimensions related to size or location can be automatically evaluated, limiting its universality.

In this paper, we propose I^{2}EBench, a comprehensive benchmark designed to automatically evaluate the performance of IIE models. I^{2}EBench exhibits three attractive characteristics: 1) Comprehensive Evaluation Dimension, 2) Human Perception Alignment, and 3) Valuable Research Insights.

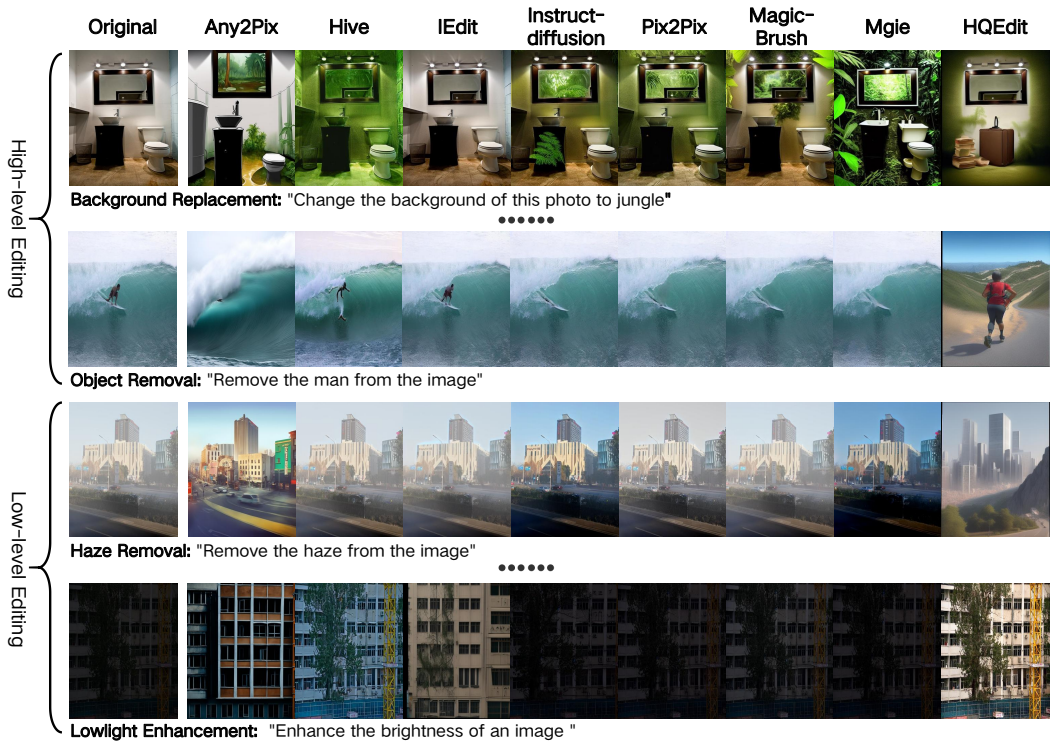


Figure 2: Visualization of the editing results on the proposed 16 evaluation dimensions using different IIE models, including InstructAny2Pix Li et al. [2023c], HIVE Zhang et al. [2023a], InstructEdit Wang et al. [2023b], InstructDiffusion Geng et al. [2023], InstructPix2Pix Brooks et al. [2023], MagicBrush Zhang et al. [2024a], MGIE Fu et al. [2024], and HQEdit Hui et al. [2024]. A detailed version can be found in supplementary materials.

First and foremost, I^2EBench offers a comprehensive evaluation dimension. These dimensions are categorized into two main types: *High-level Editing* and *Low-level Editing*. High-level editing primarily focuses on understanding instructions or editing specific areas of images, whereas low-level editing is more concerned with editing image details or the entire image. As shown in Fig. 1, both high-level and low-level editing consist of 8 fine-grained editing dimensions, which serve to demonstrate the model's proficiency in high-level and low-level editing. We meticulously collected approximately 140 images for each editing dimension and annotated each image with an original editing text instruction. To diversify the instructions, we also utilized ChatGPT Achiam et al. [2023] to enhance the description of the instructions and obtain a wider range of variations. In addition to the multi-dimensional evaluation, we also conducted a multi-category evaluation to assess the model's performance on different content categories. To achieve this, we included additional annotations for each instruction with different categories such as Animal, Object, Scenery, Plant, Human, and Global.

Second, the I^2EBench score aligns with human perception. This is accomplished by collecting scores from human annotators for the outputs generated by different IIE models, covering multiple evaluation dimensions. By conducting a comprehensive analysis of both the I^2EBench scores and the human scores, we have identified a substantial correlation between them. This discovery serves as compelling evidence, affirming that our proposed evaluation approach closely aligns with human perception.

Lastly, I^2EBench offers valuable research insights through its systematic evaluation across various dimensions and categories. The proposed I^2EBench not only facilitates a comprehensive assessment of existing models but also derives valuable insights into their respective strengths and weaknesses. These insights act as a roadmap for enhancing architecture design, refining data selection strategies, and ultimately elevating the quality of editing outcomes.

We are open-sourcing I^2EBench, including all instructions, input images, human annotations, edited images from all evaluated methods (like Fig. 2), and a simple script for evaluating the results of new IIE models. By making these resources freely available, we aim to foster fair comparisons within the field and facilitate valuable insights for community development.

2 Related Work

2.1 Instruction-based Image Editing

With the advancements in Generative Adversarial Networks (GAN)Goodfellow et al. [2014, 2020], Mao et al. [2017], Karras et al. [2019], Yoon et al. [2019], Karras et al. [2020a], Chen et al. [2018], Zhang et al. [2019] and Diffusion modelsSong et al. [2020], Ho et al. [2020], Nichol and Dhariwal [2021], Kawar et al. [2022], Austin et al. [2021], Dockhorn et al. [2022], text-to-image models Saharia et al. [2022], Rombach et al. [2022], Ramesh et al. [2021, 2022], Betker et al. [2023], Karras et al. [2019, 2020b] have made remarkable progress in recent years. As the demand for image editing continues to grow, a multitude of text-based image editing Xu et al. [2024], Kawar et al. [2023], Zhang et al. [2023b], Saund et al. [2003], Zhang et al. [2024b] models have emerged. One editing task, known as Prompt-based Image Editing (PIE) Avrahami et al. [2022], Valevski et al. [2023], Hertz et al. [2022], Dong et al. [2023], requires users to provide a target description along with the original image. The PIE model then analyzes the target description to modify the input image accordingly, generating a target image that matches the provided description. However, despite the lowered threshold for image editing, the requirement of describing the entire content of the target image in the description still poses challenges in terms of user interaction. To address this limitation, Instruction-based Image Editing (IIE) Brooks et al. [2023], Geng et al. [2023], Li et al. [2023c], Fu et al. [2024], Huang et al. [2024b] was proposed, which simplifies the user's role to providing the original image and modification instructions (*e.g.,* 'Remove the dog'). One notable implementation, InstructPix2Pix Brooks et al. [2023], introduces a large-scale dataset for instruction-based image editing. The dataset is created using a fine-tuned GPT-3 Brown et al. [2020] and image pairs generated by the Prompt-to-Prompt diffusion model Hertz et al. [2022]. Additionally, InstructPix2Pix proposes an instruction-based diffusion model for image editing based on this dataset. However, due to the automatic generation and filtering of the InstructPix2Pix dataset, concerns arise regarding its quality and potential noise. To address this, MagicBrush Zhang et al. [2024a] proposes a manually-annotated instruction-guided image editing dataset. In addition to textual instructions, InstructAny2Pix Li et al. [2023c] proposes a model that utilizes other modalities, such as audio and image, as instructions. To enhance the level of detail in instructions and improve the accuracy of editing results, MGIE Fu et al. [2024] introduces the use of Multimodal Large Language Models (MLLM) Liu et al. [2023a]. SmartEdit Achiam et al. [2023], aiming to improve the editing capabilities of IIE models in complex scenes, incorporates MLLM into the IIE model to better comprehend instructions. Despite significant progress, evaluating the editing performance of IIE models remains a crucial concern. Therefore, in this paper, we present I^2EBench, a systematic evaluation framework for these models. Our work includes an in-depth analysis of their strengths and weaknesses, offering valuable insights for the future development of IIE models.

2.2 Text-based Image Editing Benchmark

While numerous benchmarks Marino et al. [2019], Hudson and Manning [2019], Bigham et al. [2010], Lu et al. [2022], Li et al. [2023d,a,b], Yu et al. [2023], Wu et al. [2023b] have been introduced for evaluating vision-language tasks Wang et al. [2024b,a, 2022], Wu et al. [2023a], Dai et al. [2024], Hu et al. [2024], the evaluation of text-based image editing models often relies on metrics such as CLIP Score Radford et al. [2021], PSNR Korhonen and You [2012], SSIM Wang et al. [2004], and LPIPS Zhang et al. [2018]. Several existing studies have introduced benchmarks to assess the performance of image editing models. TedBench Kawar et al. [2023] presents a relatively small benchmark consisting of only 100 images and a limited set of highly common editing types. EditBench Wang et al. [2023c] is specifically designed to evaluate mask-guided image editing methods, which necessitate the availability of additional masks indicating the areas to be edited. In EditVal Basu et al. [2023], the evaluation of certain dimensions relies on manual labor, thereby limiting the reproducibility of performance. Moreover, the remaining dimensions primarily involve modifications to object size or position, lacking comprehensive coverage. While MagicBrush Zhang et al. [2024a] and Emu Edit Sheynin et al. [2023] propose test sets for evaluating editing performance,

they still rely on conventional metrics such as L1, L2, CLIP-I, DINO, and CLIP-T, which may not accurately capture the nuances of all editing types. SmartEdit Huang et al. [2024b] specifically develops a benchmark tailored for complex editing scenarios, but it does not accommodate other editing scenarios. Considering the current absence of a systematic benchmark that comprehensively evaluates the editing performance of IIE models across different editing types, we propose I^2EBench to address this gap.

3 I^2EBench

This section provides an overview of the main components of I^2EBench. In Sec. 3.1, we provide a concise introduction to the principles, definitions, and evaluation methods of 16 dimensions. Sec. 3.2 outlines the process of data annotation. Lastly, in Sec. 3.3, we present the human evaluation process to assess the correlation between the I^2EBench score and the human score. *A detailed explanation can be found in the supplementary materials.*

3.1 Evaluation Dimension

In our evaluation of the IIE model's editing quality, we have categorized it into 16 dimensions, each assessing different aspects of editing in a top-down manner. An overview of I^2EBench is presented in Fig. 1. High-level Editing Evaluation primarily focuses on assessing the model's ability to accurately understand instructions and make precise edits to local areas of the input image. This evaluation consists of 8 dimensions. Low-level Editing Evaluation, on the other hand, primarily examines global editing and detailed image processing. It also comprises 8 evaluation dimensions. Unlike previous approaches Fu et al. [2024], Zhang et al. [2023a], Geng et al. [2023] that relied on a single metric, such as CLIP score Radford et al. [2021], to evaluate editing quality for all editing types, we have developed specialized evaluation methods for each of the 16 dimensions. This approach is necessary due to the distinct goals of high-level and low-level editing.

3.1.1 High-level Editing

Evaluating editing quality in high-level dimensions poses a challenge due to the diverse goals involved, making it impractical to rely on a single metric. The advancement of Multimodal Large Language Models (MLLM) Gao et al. [2024], Chu et al. [2024], Zhu et al. [2024], Dong et al. [2024], Ma et al. [2022, 2023], Ji et al. [2022], such as GPT-4V Achiam et al. [2023], Gemini Pro Reid et al. [2024], and QWen-VL-Plus Bai et al. [2023], has significantly enhanced automated understanding of images. Therefore, to ensure precise evaluation of the editing quality of IIE models in high-level dimensions, we leverage the exceptional capabilities of the widely recognized GPT-4V model to make judgments for most high-level evaluation dimensions.

Counting. The Counting dimension pertains to instructions related to the number of objects, such as "add two apples to the image." To assess this dimension, we query GPT-4V about the number of target objects in the image and compare its response with the human-annotated answer.

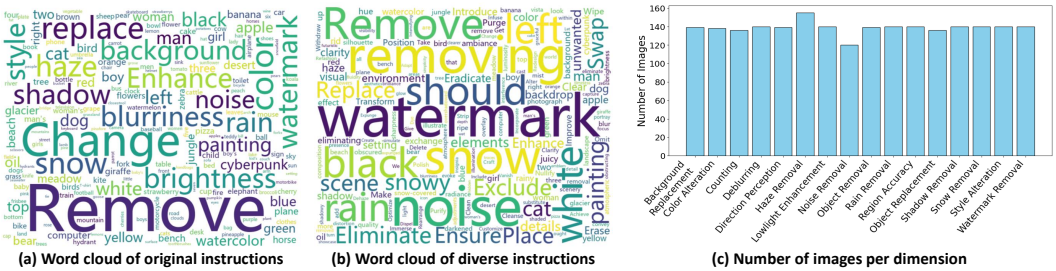
Direction Perception. The Direction Perception dimension requires the IIE model to comprehend directions provided in instructions, and accurately make edits when presented with images. We evaluate this dimension by asking GPT-4V if the target object is in the expected position.

Object Removal. The Object Removal dimension focuses on removing the target object according to the given instruction. To evaluate this dimension, we inquire whether GPT-4V identifies the presence of the target object in the image.

Object Replacement. The Object Replacement dimension aims to replace the original object with the target object as instructed. To assess this dimension, we query GPT-4V about the presence of the target object in the image.

Background Replacement. The Background Replacement dimension involves replacing the original background with the target background as specified in the instruction. To evaluate this dimension, we ask GPT-4V if the background of the image matches the textual instruction.

Color Alteration. In the Color Alteration dimension, we modify the color of the target object using instructions. To evaluate this dimension, we inquire GPT-4V about the color of the target object in the edited image.



(a) Word cloud of original instructions (b) Word cloud of diverse instructions (c) Number of images per dimension

Figure 3: Word cloud visualization (a,b) and image quantity statistics (c) of I^2EBench.

Style Alteration. The Style Alteration dimension focuses on changing the style of the image. To evaluate this dimension, we calculate the CLIP similarity Radford et al. [2021] between the edited image and "an image with $\cdots$ style".

Region Accuracy. In the editing task, we not only assess whether the target area has been edited correctly but also whether areas that should not be edited have been altered. To evaluate this dimension, we sample input images and instructions from the Object Removal, Object Replacement, and Color Alteration dimensions. We annotate the mask for the area that requires editing. Next, we fill the mask area of the images before and after editing with white and calculate SSIM Wang et al. [2004] to evaluate this dimension.

3.1.2 Low-level Editing

Unlike high-level editing, low-level editing instructions are simpler, lacking specifications regarding object size, orientation, or color. Various low-level editing tasks Wang et al. [2023a], Chen et al. [2023a], Sanghvi et al. [2023], Chen et al. [2023b], Wu et al. [2023c], Guo et al. [2023a], Kong et al. [2022] have undergone extensive development over the years, resulting in a relatively mature evaluation system. Therefore, for low-level editing, we employ the widely recognized metric, namely SSIM Wang et al. [2004], to evaluate the editing quality.

Deblurring. Deblurring encompasses the procedure of mitigating or eliminating blur from images, resulting in enhanced clarity and sharpness.

Haze Removal. Haze removal entails the elimination or reduction of atmospheric haze or fog from images, augmenting visibility and reinstating the true colors and intricate details of the scene.

Lowlight Enhancement. Lowlight enhancement refers to the process of improving the quality of images captured in low-light conditions, enhancing brightness, and reducing noise.

Noise Removal. Noise removal involves the reduction or elimination of unwanted noises in images, resulting in cleaner and more visually appealing visuals.

Rain Removal. Rain removal aims to eliminate or reduce the visual effects of raindrops or rain streaks from images, improving clarity and restoring the original appearance.

Shadow Removal. Shadow removal refers to reducing or eliminating unwanted shadows from images, enhancing visibility, and improving overall image quality.

Snow Removal. The goal of Snow Removal is to effectively reduce or eliminate snow from images.

Watermark Removal. Watermark removal involves the removal or elimination of embedded watermarks from images, restoring the original appearance without the presence of the watermark.

3.2 Human Annotation

Data Annotation. We meticulously curated approximately 140 images from publicly available datasets Lin et al. [2014], Guo et al. [2023b], Martin et al. [2001], Chen et al. [2021], Ancuti et al. [2019], Liu et al. [2021b,a], Qu et al. [2017], Nah et al. [2017], Shen et al. [2019], Wei et al. [2018] for each evaluation dimension of I^2EBench. The distribution of the image count for each dimension is illustrated in Fig. 3 (c). These images were then meticulously annotated with textual editing instructions by human annotators, namely original instructions. However, instructions provided by

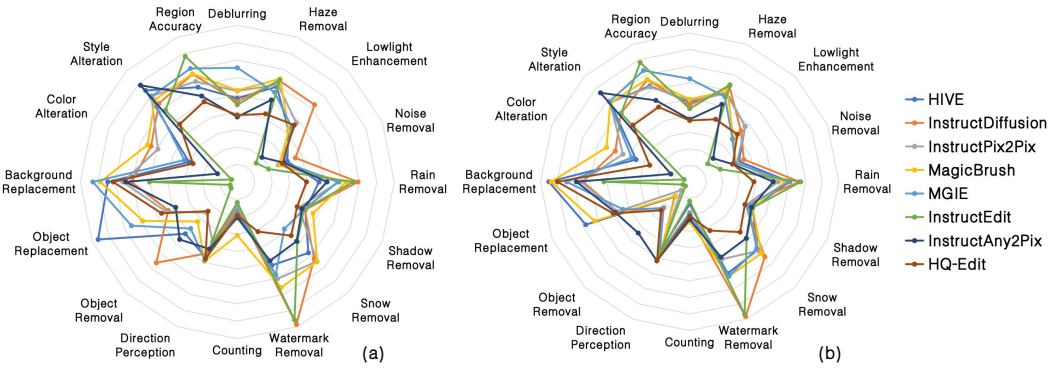


Figure 4: Comparison of radar charts for I^{2}EBench scores in different dimensions using (a) original instructions and (b) diverse instructions.

human annotators usually followed a singular sentence pattern. For instance, the prevalent sentence pattern for the object removal dimension was typically "remove $\cdots$ from the image". To foster increased diversity, we employed ChatGPT Achiam et al. [2023] to effectively rewrite the original instructions. Fig. 3 (a) and (b) present the word cloud visualizations of the original and diverse instructions, respectively. Additionally, we also annotate a category for each instruction, such as animal, object, scenery, plant, human, and global.

Evaluation Annotation. The evaluation process for I^{2}EBench encompasses two distinct categories. The first category employs conventional metrics to assess various dimensions. For style alteration dimension, we utilize the CLIP score as a standard metric, which doesn't need any additional evaluation annotations. In the second category, we utilize GPT-4V to evaluate the quality of editing. To facilitate this evaluation, we enlisted the expertise of human annotators to annotate questions specifically designed for GPT-4V, along with corresponding standard answers. For instance, let's consider the counting dimension and the instruction "Add a cat to the shoe rack". In this particular case, the annotated question provided by the human annotators is "How many cats are there on the shoe rack?", and the corresponding annotated answer is "One".

3.3 Human Evaluation

The primary objective of the human evaluation is to ascertain the correlation between human perception and the I^{2}EBench score. To achieve this, we present human evaluators with a textual instruction T, an input image V_I, and a set of edited images $\{V_1, V_2, \cdots, V_M\}$ generated by M different IIE models. The evaluators are then tasked with ranking the results based on their judgment. More specifically, we sample N images for each evaluation dimension, leading to a comprehensive collection of $N \times 16 \times 2$ edited image comparisons. Within each comparison, evaluators are presented with M edited images to assess and rank in relation to one another. We assign a human score to each model based on its ranking among the M models. Specifically, the model ranked first among the M models receives a human score of M, while the model ranked last among the M models receives a human score of 1. Additionally, the model ranked k among the M models is assigned a human score of $M - k + 1$. To determine the human score for each dimension, we calculate the average of the human scores across all samples within that dimension. Thus, the human score for each model ranges from 1 to M.

4 Experiments

Dimension Evaluation. For each image and instruction, we utilize official codes from various models for image editing. We calculate the I^{2}EBench scores following the methodology described in Sec.3.1. The I^{2}EBench scores for original and diverse instructions are presented in Fig. 4, Tab.1, and Tab. 2, respectively. Our observations reveal that no single model achieves the best performance across all evaluation dimensions. Regarding low-level editing, InstructDiffusion Geng et al. [2023] demonstrates superior results. It attains the highest scores in 4 out of 7 low-level editing evaluation dimensions when using original instructions, and 3 out of 7 when using diverse instructions. For

Table 1: I^{2}EBench evaluation results per dimension using original instructions. *Exp Min* and *Exp Max* denote the minimum and maximum values of all samples for each evaluation dimension.

Low-level Editing								
Model	**Deblurring**	**Haze Removal**	**Lowlight Enhancement**	**Noise Removal**	**Rain Removal**	**Shadow Removal**	**Snow Removal**	**Watermark Removal**
HIVE Zhang et al. [2023a]	44.25	54.89	37.61	24.59	45.47	37.61	51.49	49.99
InstructDiffusion Geng et al. [2023]	42.48	58.45	**56.61**	**28.60**	**67.20**	37.43	55.65	**85.49**
InstructPix2Pix Brooks et al. [2023]	48.03	56.15	43.32	20.11	56.64	34.19	57.59	58.12
MagicBrush Zhang et al. [2024a]	48.38	**59.46**	37.71	20.59	60.60	**41.91**	**57.81**	63.33
MGIE Fu et al. [2024]	**60.30**	51.75	39.99	23.25	56.00	36.91	34.04	55.53
InstructEdit Wang et al. [2023b]	40.77	58.85	13.83	15.40	64.44	36.88	43.45	82.68
InstructAny2Pix Li et al. [2023c]	34.34	47.27	18.03	22.89	49.94	35.84	42.97	47.28
HQ-Edit Hui et al. [2024]	35.27	39.25	41.71	22.13	38.52	33.13	38.97	29.80
Exp Min	13.79	12.66	0.09	0.79	7.38	1.05	2.18	1.34
Exp Max	91.94	92.70	89.60	77.00	96.11	89.19	89.26	96.42

High-level Editing								
Model	**Counting**	**Direction Perception**	**Object Removal**	**Object Replacement**	**Background Replacement**	**Color Alteration**	**Style Alteration**	**Region Accuracy**
HIVE Zhang et al. [2023a]	18.57	47.14	42.14	**86.43**	74.29	30.00	25.32	58.15
InstructDiffusion Geng et al. [2023]	15.00	44.29	**65.71**	42.86	60.71	53.57	21.69	66.18
InstructPix2Pix Brooks et al. [2023]	13.57	37.14	25.00	44.29	65.71	49.29	23.76	61.63
MagicBrush Zhang et al. [2024a]	**30.71**	**49.29**	32.14	58.57	78.57	**55.71**	22.78	66.34
MGIE Fu et al. [2024]	17.14	48.57	37.86	65.71	**82.86**	32.86	23.68	69.60
InstructEdit Wang et al. [2023b]	11.76	41.73	5.04	4.41	50.36	3.62	19.83	**77.08**
InstructAny2Pix Li et al. [2023c]	20.59	41.73	46.76	38.24	64.03	12.32	**26.76**	52.75
HQ-Edit Hui et al. [2024]	19.26	47.79	23.74	47.06	71.22	27.54	15.96	49.21
Exp Min	0.00	0.00	0.00	0.00	0.00	0.00	12.96	6.41
Exp Max	100.00	100.00	100.00	100.00	100.00	100.00	33.84	98.70

Table 2: I^{2}EBench evaluation results per dimension using diverse instructions.

Low-level Editing								
Model	**Deblurring**	**Haze Removal**	**Lowlight Enhancement**	**Noise Removal**	**Rain Removal**	**Shadow Removal**	**Snow Removal**	**Watermark Removal**
HIVE Zhang et al. [2023a]	44.41	54.09	42.78	25.51	58.59	36.69	51.92	57.88
InstructDiffusion Geng et al. [2023]	42.62	58.01	39.47	**28.06**	64.18	32.54	**57.30**	**85.14**
InstructPix2Pix Brooks et al. [2023]	45.24	53.52	**42.88**	24.49	51.86	32.79	52.67	48.91
MagicBrush Zhang et al. [2024a]	45.96	55.11	33.74	23.91	55.77	**36.73**	54.68	59.76
MGIE Fu et al. [2024]	**57.33**	51.61	32.96	23.49	58.27	34.07	51.02	59.64
InstructEdit Wang et al. [2023b]	40.66	**58.89**	13.92	15.81	**65.08**	36.66	43.34	83.68
InstructAny2Pix Li et al. [2023c]	34.77	47.00	18.09	22.18	48.92	36.04	43.13	47.58
HQ-Edit Hui et al. [2024]	34.11	37.95	36.76	22.38	37.60	32.17	38.45	30.83
Exp Min	6.32	3.67	0.60	0.03	7.22	1.46	3.78	2.58
Exp Max	88.17	92.69	90.34	79.29	97.03	86.27	82.24	96.39

High-level Editing								
Model	**Counting**	**Direction Perception**	**Object Removal**	**Object Replacement**	**Background Replacement**	**Color Alteration**	**Style Alteration**	**Region Accuracy**
HIVE Zhang et al. [2023a]	13.57	43.57	12.86	**67.86**	**85.00**	35.00	23.08	61.97
InstructDiffusion Geng et al. [2023]	21.43	47.86	22.14	47.14	64.29	48.57	19.96	65.92
InstructPix2Pix Brooks et al. [2023]	18.57	47.86	7.14	47.14	65.71	43.57	23.13	61.32
MagicBrush Zhang et al. [2024a]	**24.29**	45.71	12.14	62.14	83.57	**54.29**	23.08	66.21
MGIE Fu et al. [2024]	19.29	47.14	22.86	43.57	74.29	37.86	23.36	71.89
InstructEdit Wang et al. [2023b]	11.76	46.04	3.60	4.41	51.80	3.62	19.91	**77.08**
InstructAny2Pix Li et al. [2023c]	22.79	**51.80**	**43.88**	48.53	68.35	12.32	**25.93**	52.61
HQ-Edit Hui et al. [2024]	20.74	51.47	24.46	50.00	79.86	26.09	16.48	48.29
Exp Min	0.00	0.00	0.00	0.00	0.00	0.00	10.62	9.79
Exp Max	100.00	100.00	100.00	100.00	100.00	100.00	34.06	98.68

high-level editing, both MagicBrush Zhang et al. [2024a] and InstructAny2Pix Li et al. [2023c] perform impressively. MagicBrush achieves the highest scores in 3 evaluation dimensions using original instructions, while InstructAny2Pix achieves the highest scores in 3 dimensions using diverse instructions. In the deblurring dimensions, MGIE Fu et al. [2024] stands out significantly. It surpasses the second-place model by 11.92 when using original instructions and by 11.37 when using diverse instructions.

Human Evaluation. We ranked different models based on their I^{2}EBench scores and computed I^{2}EBench rank scores using the methodology described in Sec. 3.3. Given that both I^{2}EBench rank scores and human scores range from 1 to 8, a direct comparison can be made between them. Therefore, we conducted correlation analyses and visually presented the results in Fig. 5. Significant positive correlations were observed between the I^{2}EBench rank score and the human score across all

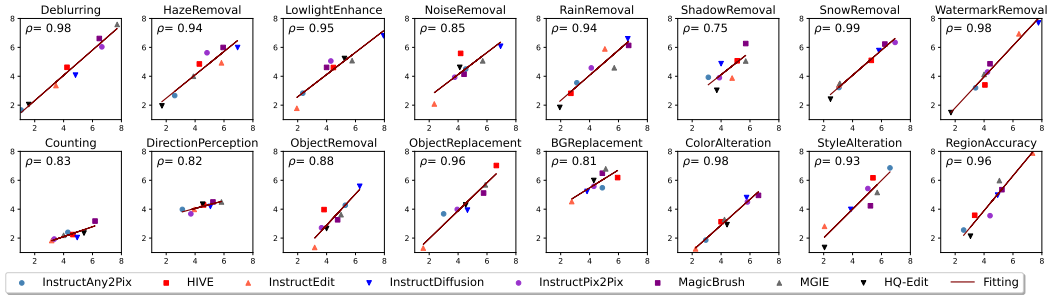


Figure 5: Alignment between I^{2}EBench rank scores (Y-axis) and human scores (X-axis).

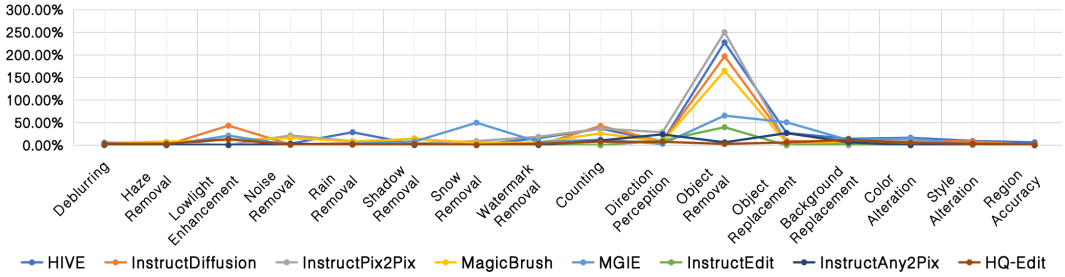


Figure 6: I^{2}EBench change rate using original instructions and diverse instructions.

dimensions. These findings offer strong evidence supporting the alignment between our proposed benchmark and human perception.

5 Insights

The editing ability across different dimensions is not robust: Our observations indicate that no single model excels in all evaluation dimensions. This implies that different IIE models have varying strengths in terms of their editing abilities across different dimensions. Thus, it is crucial to acknowledge this limitation and focus on developing an IIE model that demonstrates consistent and competent performance across all dimensions. *Future research efforts should prioritize the creation of a robust and versatile IIE model that can effectively handle a wide range of editing tasks across diverse dimensions.*

The editing ability of different instructions is not robust: To evaluate the robustness of editing models when provided with different instructions, we propose a metric called I^{2}EBench change rate. This metric is defined as follows:

$$S^i = \frac{|S_o^i - S_d^i|}{\text{MIN}(S_o^i, S_d^i)}, \tag{1}$$

where S_o^i and S_d^i represent the I^{2}EBench scores of the i-th evaluation dimension when using original and diverse instructions, respectively. The value of S^i indicates the I^{2}EBench change rate for the i-th evaluation dimension. As illustrated in Fig. 6, when it comes to the object removal dimension, InstructPix2PixBrooks et al. [2023], HIVE Zhang et al. [2023a], InstructionDiffusion Geng et al. [2023], and MagicBrush Zhang et al. [2024a] exhibit significant fluctuations in their performance using different instructions. On the other hand, the remaining models demonstrate relatively stable performance across different instructions. One notable distinction between these two categories of models is that the latter employs LLM Achiam et al. [2023], Touvron et al. [2023] or MLLM Liu et al. [2023b,a] to comprehend instructions, which enhances their resilience to variations in instructions. *Given the unpredictable and diverse nature of user editing instructions, it is crucial to develop an editing model that can effectively handle instructions with varying levels of complexity.*

The editing ability for different categories is not robust: As illustrated in Fig. 7, we have observed distinct variations in the performance of different categories. Notably, the "Scenery" and "Global"

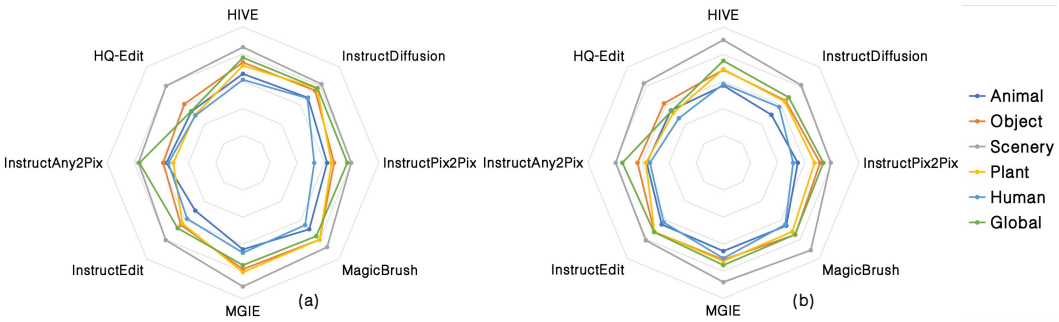


Figure 7: Comparison of radar charts for I^{2}EBench scores in different categories using (a) original instructions and (b) diverse instructions. The scores of all dimensions are normalized and averaged.

categories consistently demonstrate superior performance compared to the other categories across all the IIE models we evaluated. This discrepancy can be attributed to the inherent inclination of the "Scenery" and "Global" categories towards global editing, which diminishes the necessity for precise target object localization. *Given these findings, it is crucial to prioritize the simultaneous consideration of various editing content in future research endeavors.*

6 Conclusions

In this paper, we present I^{2}EBench, a comprehensive benchmark specifically designed for instruction-based image editing (IIE). Our benchmark includes a substantial dataset of over 2000+ images and more than 4000+ instructions, covering 16 distinct evaluation dimensions. To evaluate the effectiveness of I^{2}EBench, we conduct experiments using 8 open-source IIE models. Additionally, we complement these experiments with meticulous human evaluations to establish the correlation between I^{2}EBench scores and human perception. Based on the observations derived from I^{2}EBench, we provide valuable insights and recommendations for advancing IIE models. We hope the proposed I^{2}EBench to serve as an indispensable asset, playing a pivotal role in fostering the advancement of IIE models and assessing their efficacy.

Acknowledge

This work was supported by National Key R&D Program of China (No.2023YFB4502804), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U22B2051, No. U21B2037, No. 62072389, No. 62302411), the Natural Science Foundation of Fujian Province of China (No.2021J06003), and China Postdoctoral Science Foundation (No. 2023M732948).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Codruta O Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In *2019 IEEE international conference on image processing (ICIP)*, pages 1014–1018. IEEE, 2019.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Samyadeep Basu, Mehrdad Saberi, Shweta Bhardwaj, Atoosa Malemir Chegini, Daniela Massiceti, Maziar Sanjabi, Shell Xu Hu, and Soheil Feizi. Editval: Benchmarking diffusion based text-guided image editing methods. *arXiv preprint arXiv:2310.02426*, 2023.

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.

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.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Haoyu Chen, Jinjin Gu, Yihao Liu, Salma Abdel Magid, Chao Dong, Qiong Wang, Hanspeter Pfister, and Lei Zhu. Masked image training for generalizable deep image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1703, 2023a.

Wei-Ting Chen, Hao-Yu Fang, Cheng-Lin Hsieh, Cheng-Che Tsai, I Chen, Jian-Jiun Ding, Sy-Yen Kuo, et al. All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4196–4205, 2021.

Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5896–5905, 2023b.

Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. Cartoongan: Generative adversarial networks for photo cartoonization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9465–9474, 2018.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Genie: Higher-order denoising diffusion solvers. *Advances in Neural Information Processing Systems*, 35:30150–30166, 2022.

Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7430–7440, 2023.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Forty-first International Conference on Machine Learning*, 2024a.

Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing, 2024b.

Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024c.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *International Conference on Learning Representations*, 2024.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024.

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, et al. Instructdiffusion: A generalist modeling interface for vision tasks. *arXiv preprint arXiv:2309.03895*, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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

Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14049–14058, 2023a.

Yun Guo, Xueyao Xiao, Yi Chang, Shumin Deng, and Luxin Yan. From sky to the ground: A large-scale benchmark and simple baseline towards real rain removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12097–12107, 2023b.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024.

Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey. *arXiv preprint arXiv:2402.17525*, 2024a.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024.

Jiayi Ji, Yiwei Ma, Xiaoshuai Sun, Yiyi Zhou, Yongjian Wu, and Rongrong Ji. Knowing what to learn: a metric-oriented focal mechanism for image captioning. *IEEE Transactions on Image Processing*, 31:4321–4335, 2022.

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.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020a.

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

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023.

Xiangtao Kong, Xina Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Reflash dropout in image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6002–6012, 2022.

Jari Korhonen and Junyong You. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth International Workshop on Quality of Multimedia Experience*, pages 37–38. IEEE, 2012.

Vladimir Kulikov, Shahar Yadin, Matan Kleiner, and Tomer Michaeli. Sinddm: A single image denoising diffusion model. In *International Conference on Machine Learning (ICML)*, pages 17920–17930. PMLR, 2023.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*, 2023b.

Shufan Li, Harkanwar Singh, and Aditya Grover. Instructany2pix: Flexible visual editing via multimodal instruction following. *arXiv preprint arXiv:2312.06738*, 2023c.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023d.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b.

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

Ye Liu, Lei Zhu, Shunda Pei, Huazhu Fu, Jing Qin, Qing Zhang, Liang Wan, and Wei Feng. From synthetic to real: Image dehazing collaborating with unlabeled real data. In *Proceedings of the 29th ACM international conference on multimedia*, pages 50–58, 2021b.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022.

Yiwei Ma, Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Towards local visual modeling for image captioning. *Pattern Recognition*, 138:109420, 2023.

Yiwei Ma, Zhibin Wang, Xiaoshuai Sun, Weihuang Lin, Qiang Zhou, Jiayi Ji, and Rongrong Ji. Inf-llava: Dual-perspective perception for high-resolution multimodal large language model. *arXiv preprint arXiv:2407.16198*, 2024.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4067–4075, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pages 8748–8763. PMLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Yash Sanghvi, Zhiyuan Mao, and Stanley H Chan. Structured kernel estimation for photon-limited deconvolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9863–9872, 2023.

Eric Saund, David Fleet, Daniel Larner, and James Mahoney. Perceptually-supported image editing of text and graphics. In *Proceedings of the 16th annual ACM symposium on User interface software and technology*, pages 183–192, 2003.

Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5572–5581, 2019.

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.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning (ICML)*, pages 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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

Dani Valevski, Matan Kalman, Eyal Molad, Eyal Segalis, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning a diffusion model on a single image. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Hang Wang, Xuanhong Chen, Bingbing Ni, Yutian Liu, and Jinfan Liu. Omni aggregation networks for lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22387, 2023a.

Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions. *arXiv preprint arXiv:2305.18047*, 2023b.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18359–18369, 2023c.

Yabing Wang, Jianfeng Dong, Tianxiang Liang, Minsong Zhang, Rui Cai, and Xun Wang. Cross-lingual cross-modal retrieval with noise-robust learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 422–433, 2022.

Yabing Wang, Fan Wang, Jianfeng Dong, and Hao Luo. Cl2cm: Improving cross-lingual cross-modal retrieval via cross-lingual knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5651–5659, 2024a.

Yabing Wang, Shuhui Wang, Hao Luo, Jianfeng Dong, Fan Wang, Meng Han, Xun Wang, and Meng Wang. Dual-view curricular optimal transport for cross-lingual cross-modal retrieval. *IEEE Transactions on Image Processing*, 33:1522–1533, 2024b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML)*, pages 681–688, 2011.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023a.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023b.

Rui-Qi Wu, Zheng-Peng Duan, Chun-Le Guo, Zhi Chai, and Chongyi Li. Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22282–22291, 2023c.

Sihan Xu, Ziqiao Ma, Yidong Huang, Honglak Lee, and Joyce Chai. Cyclenet: Rethinking cycle consistency in text-guided diffusion for image manipulation. *Advances in Neural Information Processing Systems*, 36, 2024.

Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

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, 2024a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

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*, 2023a.

Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6027–6037, 2023b.

Zhongping Zhang, Jian Zheng, Zhiyuan Fang, and Bryan A Plummer. Text-to-image editing by image information removal. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5232–5241, 2024b.

Dongsheng Zhu, Xunzhu Tang, Weidong Han, Jinghui Lu, Yukun Zhao, Guoliang Xing, Junfeng Wang, and Dawei Yin. Vislinginstruct: Elevating zero-shot learning in multi-modal language models with autonomous instruction optimization. *arXiv preprint arXiv:2402.07398*, 2024.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The last paragraph of the abstract and introduction explains the contribution and scope of this paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The limitation section is included in the appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when the image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results have been fully proven.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in the appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: This paper provides sufficient details to reproduce the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code and dataset are included in supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes] ,

   Justification: All replicates of IIE models are sourced from official settings and code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in the appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Our goal is to propose a benchmark, so there is no need for Experiment Statistical Signature.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: We did not train the IIE model, and all checkpoints are sourced from the official code, so there is no need to report them.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have made appropriate citations in the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The proposed dataset are included in Supplementary materials.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: A screenshot of the human evaluation questionnaire can be found in the appendix.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: Our experiment meets the requirements.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.