EditInspector: A Benchmark for Evaluation of Text-Guided Image Edits

Anonymous ACL submission

Abstract

Text-guided image editing, fueled by recent advancements in generative AI, is becoming increasingly widespread. This trend highlights the need for a comprehensive framework to verify text-guided edits and assess their quality. To address this need, we introduce EditInspector, a novel benchmark for evaluation of text-guided image edits, based on human annotations collected using an extensive template for edit verification. We leverage EditInspector to evaluate the performance of state-of-the-art (SoTA) vision and language models in assessing edits across various dimensions, including accuracy, artifact detection, visual quality, seamless integration with the image scene, adherence to common sense, and the ability to describe edit-induced changes. Our findings indicate that current models struggle to evaluate edits comprehensively and frequently hallucinate when describing the changes. To address these challenges, we propose two novel methods that outperform SoTA models in both artifact detection and difference caption generation.

1 Introduction

The ability to create and modify images is vital in fields such as social media, marketing, and graphic design. Recent advancements in generative AI have greatly democratized this ability. In particular, natural language enables high-quality, customized visual content creation with minimal effort.

Text-guided editing models require a source image and instruction (Kawar et al., 2023; Zhang et al., 2022; Brooks et al., 2023; Wu et al., 2023b; Zhang et al., 2024b), sometimes allowing multi-turn editing (Sheynin et al., 2023; He et al., 2024; Wu et al., 2023a; Cui et al., 2023). For more precise spatial control a user might provide the source image, a mask, and a text prompt specifying changes for the masked area (Avrahami et al., 2022; Nichol et al., 2022; Couairon et al., 2022; Wang et al., 2023; Zhang et al., 2024a). Extensive human evaluations showed that mask-based text-guided editing produces superior results compared to mask-free editing (Wang et al., 2023; Zhang et al., 2024a).

Despite these advancements, evaluating the quality and accuracy of edits remains challenging, as demonstrated in Figure 1. Current methods often focus on whether the edited object matches the requested attributes (Wang et al., 2023) or use ranking scores for accuracy (Zhang et al., 2024a). However, they overlook pain points such as unintended artifacts, misalignment with user expectation, visual quality, and adherence to common sense. For example, in Figure 2, the edit changes teardrops to stars as instructed, but unintentionally adds a line and alters the wall's appearance.

To address these challenges, we propose EditInspector, a comprehensive benchmark for *assessing evaluators of text-guided image edits* (Section 2). EditInspector examines edits across five dimensions: (1) whether the edit accurately follows the instructions and aligns with user expectations; (2) introduction of unintended artifacts; (3) technical quality (low resolution, blur, etc.); (4) the accuracy of a description of the main difference; and (5) the accuracy of a detailed listing of the differences between the original and the edited images.

We begin by creating a human evaluation framework, shown in Figure 2, that assesses edits based on the dimensions outlined above (Section 2.1). Using this framework, we collected human annotations as edit inspectors through crowdsourcing, evaluating 783 edits from the MagicBrush (Zhang et al., 2024a) test set of 1,053 edits, to introduce the EditInspector benchmark (Section 2.2).

We then evaluate state-of-the-art vision and language models (VLMs) as edit inspectors on the EditInspector benchmark, comparing their performance with human annotations, as shown in Figure 1. The results show that all models perform poorly across all tasks, with accuracy hovering

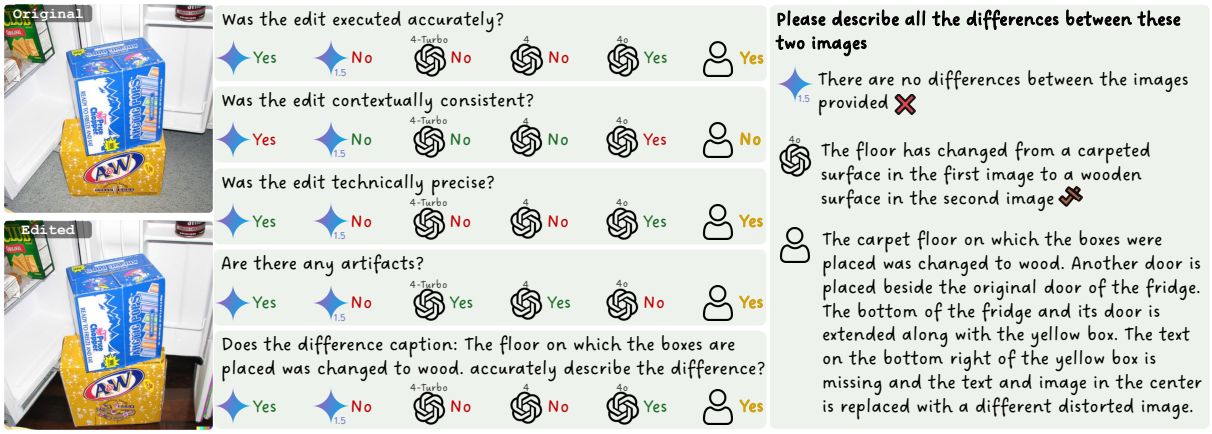

Figure 1: The assessments for the edit "Let the floor be made of wood" vary across different models, with 2–3 models answering each question correctly. Gemini 1.5 failed to detect any differences between the images, while GPT-4o successfully identified only the main difference. See Appendix A.5 for full-size prompts.

around random chance (Section 3.3.1). Gemini-1.5 (Gemini Team, 2024) emerged as the top performer for the edit inspector questions, achieving 70.3% accuracy in the edit accuracy question. We evaluate models' ability to generate a summary of the main change and a detailed list of all differences as an upper-bound test of edit accuracy, artifact detection, and visual quality. In this task, GPT-4o achieved 39% accuracy in describing the main difference but detected only 12% of all differences, with only 40% aligning with human annotations, highlighting significant hallucinations. (Section 3.3.2).

We tackle the challenges of artifact detection and difference caption generation with two methods. First, we developed a zero-shot pipeline using Gemini as the visual backbone to generate instruction-grounded difference captions and metadata (Section 4.1). The pipeline analyzes image captions at three zoom levels around the edit area and outputs a difference caption, achieving 75% accuracy in describing the main difference, compared to 39% by the best SoTA model. Second, we introduced a novel artifact detection method that achieves 64% accuracy by analyzing object segmentation probabilities around the edited area (Section 4.2).

Finally, we introduce an end-to-end fine-tuned model that rivals much larger models, delivering competitive SoTA performance while reducing computational costs (Section 5). To train our model we use two augmentation methods to generate 31,059 training instances. The first method creates negative examples with objects closely resembling the original (Section 5.1), and the second reverses the edit direction, e.g., by changing an "Add" edit to a "Remove" edit (Section 5.2).

In summary, our main contributions are: (1) A comprehensive framework for image edit evaluation, and the EditInspector benchmark, which we release for future work and future model assessment; (2) A thorough evaluation of SoTA VLMs as edit inspectors, showing that, across all aspects, none can effectively assess edits; (3) Two new methods outperforming SoTA models for artifact detection and difference caption generation; and, (4) An end-to-end fine-tuned model that rivals much larger models in performance.

2 EditInspector Dataset

Our goal is to develop a dataset and framework for image editing verification that offers a comprehensive evaluation of edits, addressing overlooked pain points like unintended artifacts, instruction inconsistencies, scene misalignment, and technical flaws. To achieve this, we introduced the human evaluation framework in Section 2.1 and annotated 783 MagicBrush edits using it to create our benchmark in Section 2.2. The statistics and analysis of our benchmark are presented in Section 2.3.

2.1 Human Evaluation Framework

Our motive was to develop a comprehensive framework that evaluates multiple aspects of image editing. We tested and refined templates and questions using internal and crowdsourced feedback, resulting in the framework shown in Figure 2.

The evaluation begins with *Accuracy Level*, where annotators assess whether the edit follows the instruction and meets user expectations. If it fully follows the instruction, annotators select *Accurate* or *Accurate, But Unexpected* if it deviates

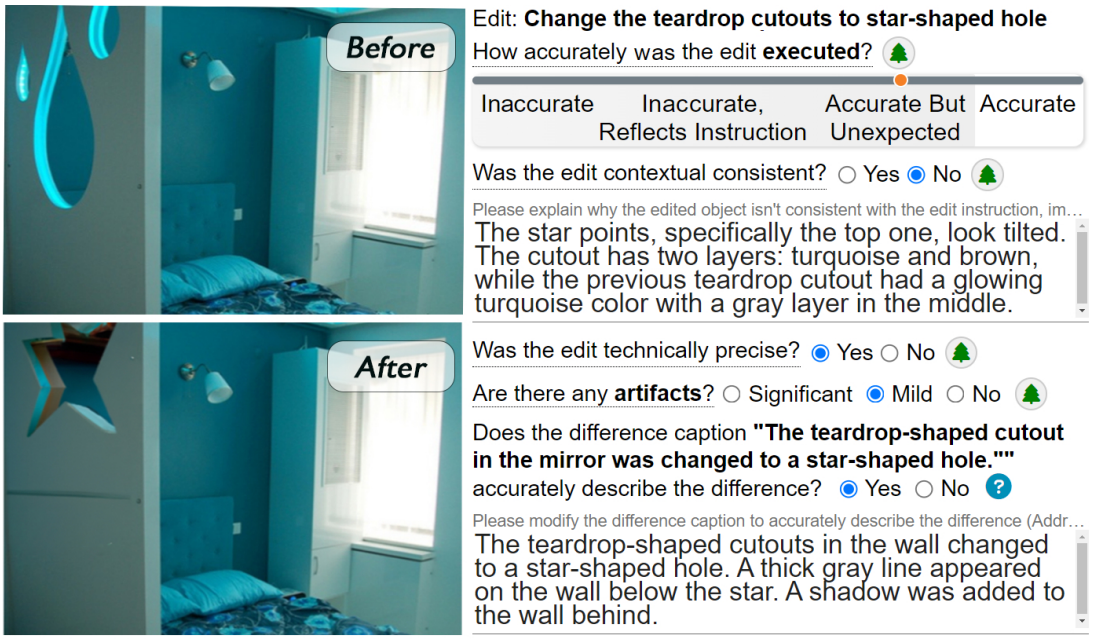

Figure 2: This is an example of our annotation user interface. The edit appears to be accurately executed but includes unexpected elements, such as differences in the door layers and a tilted star edge. There are mild artifacts, including a shadow behind the wall and a thick gray line beneath the star cutout. Clicking the tree icons opens decision trees that help annotators follow the evaluation guidelines (See Appendix A.14).

from expectations. For partial adherence to the instruction, they select *Inaccurate, Reflects Instruction*, and for no adherence, *Inaccurate*.

For any selection other than *Accurate*, annotators are asked to explain under *Contextual Consistency* how the edit failed to meet expectations or align with the instruction, image scene, or common sense. Under the *Technical Precision* question annotators comment on pixel-level details like resolution, blurriness, and smoothness.

For example, in Figure 2 a teardrop cutout was changed to a star-shaped hole, but all annotators marked it as "*Accurate, But Unexpected*" due to the tilted star edge and the unexpected material appearance, as seen in *Contextual Consistency* feedback.

Next, the *Artifacts* evaluation involves annotators identifying any unintended distortions or anomalies in the edit. Artifacts are classified into two levels: *Significant* or *Mild*, based on their severity. In example in Figure 2, two *Mild* artifacts are present: an unintended shadow and an extra line beneath the star-shaped hole.

Finally, to collect a difference caption describing all differences between the edited images as an upper-bound evaluation of the edit, we start with an automatically generated one that describes the main difference (Section 4.1). Humans then review it, either accepting or correcting it, and expand it to include additional differences if artifacts are present, as shown in Figure 2.

2.2 Human Annotation

We employed Amazon Mechanical Turk (AMT) to evaluate image edits using human annotators, as shown in Figure 2, with three annotations per edit. Quality annotators were selected through a paid qualification test, and multiple steps were taken to ensure the instructions were clear and accessible in the UI (See Appendices A.4 and A.15).

2.3 Human Evaluation Analysis

Full annotation distribution is presented in Table 1. Despite the task's subjectivity, majority agreement averaged 80% to 86%, compared to random chance of 25% for Accuracy and 33% for Artifacts. Majority agreement hit 96% for Accuracy and 97% for Artifacts. Full agreement among all annotators was achieved for 42% to 57% of edits. In 85% of examples, the edit reflected the instruction ("Accurate" or "Accurate, But Unexpected"), while 38% of edits contained significant artifacts.

The edit types, derived from metadata in Section 4.1, were distributed: Add 35.8%, Change Attribute 21.6%, Remove 7.3%, and Replace 31.3%.

Figure 3 shows the percentage of issues reported by annotators in the Contextual Consistency and Technical Precision feedback, with resolution

Category	Statistics (%)	
Accuracy Level	Accurate: 8% Accurate Unexpected: 77%	Inaccurate: 6% Inaccurate Reflects: 4%
Artifacts Level	Significant: 38%	Mild: 57% No Artifact: 2%
Technical Precision	Yes: 69%	No: 31%
Visual Consistency	Yes: 18%	No: 82%
Diff Caption Accuracy	Yes: 60%	No: 40%

Table 1: Distribution of annotation values across categories. In 85% of examples, the edit reflected the instruction ("Accurate" or "Accurate, But Unexpected"), while 38% of edits contained significant artifacts.

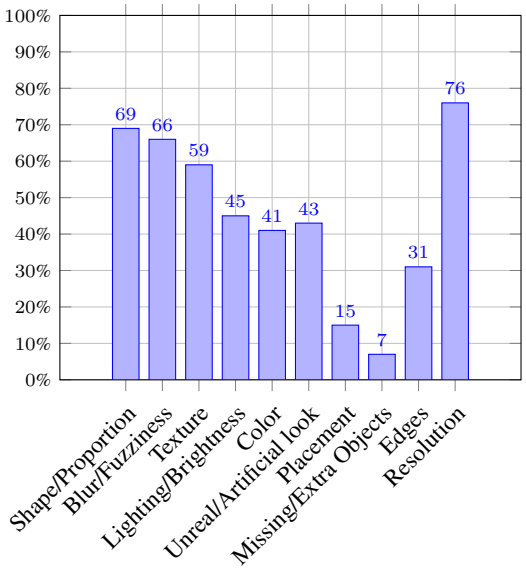

Figure 3: Frequency of issues identified by human annotators in the Contextual Consistency and Technical Precision textual feedback. Shape/Proportion concerns being particularly prominent.

and shape/proportion concerns being particularly prominent. See Appendix A.7 for a full overview.

3 Auto-Evaluation

Using the EditInspector benchmark, we evaluate the ability of SoTA VLMs to serve as edit inspectors. The evaluation consists of two components: the first assesses the models' ability to verify edit accuracy and alignment with user expectations, while the second serves as an upper-bound test, examining their ability to generate captions that describe the main differences and all differences, including unintended artifacts (Section 3.3.2).

3.1 Models

We evaluate GPT-4, GPT-4o, GPT-4-turbo (OpenAI, 2024), Gemini-Pro-Vision (Gemini Team, 2023), and Gemini-Pro-1.5 (Gemini Team, 2024) on all tasks using their latest versions as of August 2024 (Section A.10). We prioritized prompts that best conveyed user instructions and improved overall performance (See Appendix A.5).

3.2 Auto-Evaluation Setup

Edit Inspector questions. Preliminary experiments revealed that models struggled to handle multiple categories, especially in detecting mild artifacts. To enhance clarity and relevance, we simplified the categorization by replacing multiple-choice questions with binary outcome questions. For the accuracy question, both "Accurate" and "Accurate But Unexpected" were grouped under "Accurate," while in the artifacts question, only "Significant Artifacts" were counted as artifacts.

Difference Caption Generation. Traditional caption metrics (BLEU, METEOR, ROUGE, CIDEr) rely on N-gram overlaps but fail to distinguish edited objects, penalize stylistic variations, ignore edit sequences, and miss semantic misalignments. As shown in Table 2, these limitations lead to misleadingly high scores for incorrect captions. Section A.1 provides further examples and analysis.

To address these limitations, we propose two novel evaluation metrics tailored for **all differences caption** comparisons: Model Precision (MP) and Hallucination Rate (HR). MP is the percentage of human-annotated differences matching model-detected ones, while HR is the percentage of model-detected differences that do not correspond to any human-annotated differences.

We calculate these metrics by generating Difference Triplets (DTs) with the source object, target object, and action type for each change in the model and human captions. The two resulting sets of DTs are then used to compute MP and HR. A match between two DTs is determined if the edit action types are identical, and the source and target objects are similar, as evaluated by GPT-4o. The similarity check between source and target objects is relaxed, allowing matches for objects with different attributes. A stricter check would have caused models to fail completely.

In addition, we introduced MP_{soft} and HR_{soft}, which count DT matches also in case of a reversed

Example	Metrics
Ground Truth Caption: The main difference is the first image has a blue vase, and the second image has a brown vase. **Generated Caption:** The main difference is the first image has a squirrel, and the second image does not.	**MP**: 0 BL: 0.68 RO: 0.81 ME: 0.78
Ground Truth Caption: A brown squirrel was added to the image. **Generated Caption:** The difference between the two images is that the first image has a blue vase. The second image has a blue vase and a squirrel next to it.	**MP**: 1 BL: 0.55 RO: 0.60 ME: 0.57
Ground Truth Caption: In the first image, the tree was removed, and new flowerbed was added. **Generated Caption:** In the first image, the flowerbed was removed, and new tree was added.	**MP**: 0 BL: 0.73 RO: 0.79 ME: 0.76

Table 2: Comparison of traditional linguistic metrics (BLEU, ROUGE, METEOR) against our proposed evaluation metric (MP). The first example shows high scores despite missing the edited object. The second penalizes correct but longer captions. The third fails to detect reversed edits, while our metric captures these issues accurately.

source and target object match, offering a more comprehensive analysis of model performance. See Section A.2 for mathematical formulations of the metrics, and Section A.3 for an intuitive example.

We evaluate the model's **main difference caption** by comparing it to the main difference extracted from the human-provided difference caption, which describes all of the edit's differences. GPT-4 is used to assess whether the main model-identified difference matches the human one. Extracting the main difference is not complex, as the main change is mentioned first in 95% of cases.

3.3 Auto-Evaluation Results

3.3.1 Edit Inspector Questions Results

The results for the Yes/No questions are presented in Table 3. Gemini-1.5 achieved the highest score on all questions except 'Contextual Consistency', where all models performed poorly. Below, we summarize our main observations from these results.

Struggling with Inaccurate Edits and Artifact Classification. Detection of inaccurate edits was challenging, with most models correctly classifying only 0-25%, except GPT-4o (47%). All models mistakenly predicted edits as visually consistent, with precision scores between 0-22.3%. Differentiating artifacts from non-artifacts was also challenging. While GPT-4o had the highest accuracy (65.7%) it missed many artifacts with low recall (52.7%). All models frequently misclassified non-artifacts (18–30%), with Gemini misclassify 72%.

Assessing the accuracy of inconsistent edits is challenging. There is a strong conditional dependency between the edit accuracy and contextual consistency questions. A discrepancy up to 40% was observed in the accuracy question when edits lacked contextual consistency. Conversely, models had difficulty with the contextual consistency question in accurate edits, with a 23% drop in performance. This dependency was also present (up to 12%) between the caption accuracy and contextual consistency questions.

Remove edits are challenging for all models, except GPT-4o. While Gemini 1.5, GPT-4, and GPT-4-turbo struggled with 'Remove' edits, showing accuracy gaps of 36-65% in both edit and caption accuracy, GPT-4o excelled with 91% accuracy, making it the only model to handle these edits well.

Alongside Yes/No questions, we assessed models' feedback on Contextual Consistency and Technical Precision, finding it misaligned with human feedback in most cases (see Appendix A.6).

3.3.2 Difference Caption Generation Results

Main Differences Captions: Table 3 shows the percentage of instances where the model-identified main difference matched the human-reported one, with GPT-4o leading at 39% accuracy. Across all models, performance improved by up to 98% when the edit was accurate, a trend also seen in generating complete difference captions. 'Remove' edits had the lowest performance, with accuracy dropping by up to 50% compared to the best-performing 'Replace' edits.

All Differences Captions: Table 3 shows that GPT-4o achieves the highest Model Precision (MP) at 12% and the lowest Hallucination Rate (HR) at 60%, along with notable improvements in soft metrics, suggesting confusion between source and target objects. Overall performance remains suboptimal, as model predictions often misalign with human annotations. On average, models describe 1-2.5 differences per image, whereas human annotators identified six differences on average. This gap highlights models' difficulty in capturing subtle differences and their tendency to overlook details or introduce hallucinated changes.

Additionally, we observed that models tend to hallucinate less where the edits are accurately performed, leading to a 22% improvement in HR and a threefold increase in MP across all models.

Models vary significantly in predicting no dif-

	Gemini	Gemini-1.5	GPT-4	GPT-4o	GPT-4 Turbo	LLaVA	LLaVA (Supervied)
Edit Inspectors Questions							
Accuracy	49.9%	**70.3%**	67.3%	67.8%	66.9%	58.9%	67.2%
Contextual Consistency	50.4%	51.1%	50.4%	**55.7%**	48.2%	52.0%	-
Technical Precision	49.9%	**53.7%**	46.3%	45%	50.7%	49.9%	-
Artifacts	49.4%	58.5%	50.7%	**65.7%**	52.8%	47.6%	51.7%
Difference Caption Accuracy	53.9%	**66.3%**	63.9%	64.3%	64%	50.0%	54.5%
Differences Caption Generation							
Main Difference	31%	31%	27%	**39.0%**	24%	8%	10%
MP	-	8%	8%	**12%**	8%	-	-
MP_{soft}	-	9%	10%	**14%**	9%	-	-
HR	-	67%	78%	**60%**	75%	-	-
HR_{soft}	-	65%	75%	**56%**	72%	-	-
Avg. Diff	-	1	2.5	1.8	1.5	-	-
No Diffs	-	24%	0.7%	0.3%	6%	-	-

Table 3: Combined performance on Edit Inspectors questions, and the Difference Caption Generation task. Gemini-1.5 model demonstrates the best performance in Edit Inspectors questions, achieving the highest or second-highest scores across all questions. GPT-4o achieves the highest precision in predicting differences, with the lowest hallucination rate and a relatively high average number of detected differences. Avg. Diff indicates the average number of differences detected per edit, while No Diffs represents the percentage of edits where no differences were predicted. Human annotators identified an average of 6 differences per edit. The main difference row reports the percentage of predicted main difference captions correctly describing the main difference. The LLaVA (Supervised) column presents the performance of the finetuned model; see Section 5.3 for further analysis.

ferences between images. For example, Gemini-1.5 predicts no differences in 24% of the examples, compared to only 0.3% for GPT-4o. Gemini-1.5's higher rate of "no difference" predictions lowers its HR but causes it to identify fewer differences than GPT-4o, which detects 80% more differences while keeping a lower HR. When the edit is contextually consistent, models predict no differences 2 to 3 times more often, suggesting they are more sensitive to semantic flaws then pixel-level ones.

Models struggle with Remove edits while excelling in Add edits. All models perform best on Add edits and worst on Remove edits, with Model Precision (MP) differing by up to 2.7x. The Hallucination Rate (HR) for Remove edits is significantly worse, increased by 50% compared to Add edits.

Models are sensitive to scene complexity (i.e., the number of objects). Figure 6 in the Appendix shows that as the number of objects increases, all models exhibit declining precision and rising hallucination rates. GPT-4 and GPT-4-turbo, in particular, struggle more with complex scenes, showing sharp increases in hallucinations. While Gemini-1.5 and GPT-4o also degrade, their decline is less steep. This trend was not observed in the Edit Inspector questions (Yes/No questions).

4 New Methods

To tackle the challenges models face in generating accurate difference captions and detecting unintended artifacts, we developed a zero-shot pipeline for producing detailed, instruction-grounded captions (Section 4.1) and an artifact detection method using segmentation model probabilities (Section 4.2). Our methods are competitive with the best models, and in the main difference generation task outperform them by 36% margin.

4.1 Difference Caption Pipeline

Our pipeline generates detailed, instruction-grounded difference captions and rich metadata by selecting image captions of the edited object area that align with the edit instructions. **It achieves 75% accuracy in describing the main edit, surpassing GPT-4o's 39% accuracy.**

The pipeline process involves extracting image captions at three zoom levels around the edit area for both the source and target images. We then select the captions that best match the edit instructions, measured by the number of shared nouns or their synonyms using WordNet (Fellbaum, 1998). Using these grounded captions and the edit instruction, we employ a one-shot prompt with GPT-4 (OpenAI, 2024) to generate a detailed difference caption with metadata, as shown in Figure 4.

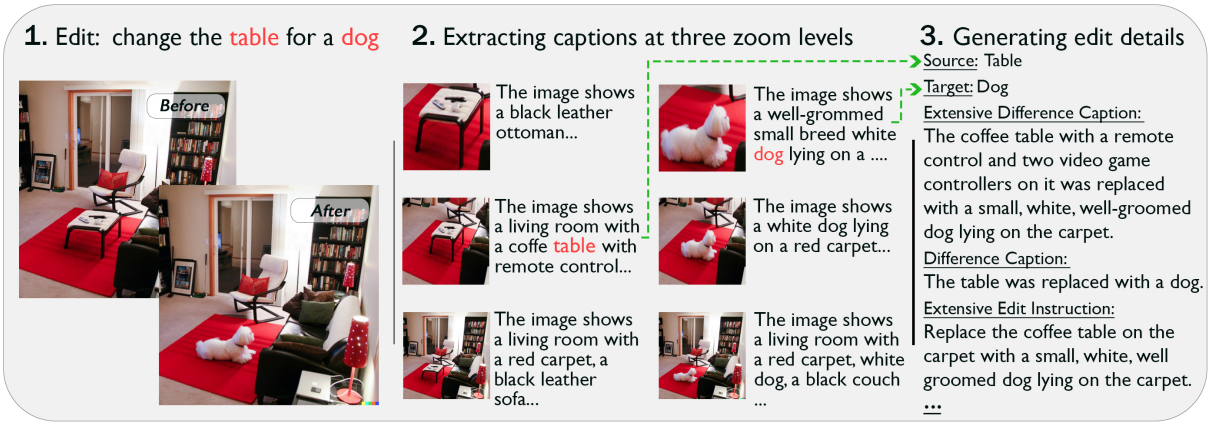

Figure 4: Example of our pipeline generating an instruction-grounded difference caption with rich metadata. Edit images are split into three zoom levels, with Gemini extracting and prioritizing captions to generate the metadata.

We found this method most effective for generating a main difference caption. Other methods, such as asking object-specific questions or requesting long image descriptions, often resulted in significant hallucinations and incorrect or biased descriptions. This issue persisted with different visual backbones, such as GPT-4 (OpenAI, 2024), LLAVA 1.5 (Liu et al., 2024), etc. Integrating human instructions with edited area descriptions allows for information as seen in Figures 16, 22.

4.2 Artifact Detection

We developed two artifact detection methods using the extracted metadata from our pipeline. The first method uses the Detic model (Zhou et al., 2022) to analyze the segmentation probability of each object intersected by the edit mask. A drop of the probability score by more than 4% as a result of the edit is considered an artifact.

The second method identifies elements that intersect with the mask area, have disappeared from the image, and do not overlap with the edited object's bounding box. This often occurs when the mask is large, but the edited object is small.

Combined, **our methods achieve 64% balanced accuracy in detecting "Significant" artifacts, competitive only with GPT-4o scoring 65.7%.** Figure 5 shows the first artifact detection method. If the small car intersecting with the inpainting area had been unintentionally removed, it would illustrate the second method.

An oracle that combines the optimal predictions from GPT-4o and our artifact detection method reaches a score of 86.8% with 100% precision. **This indicates that our artifact method and GPT-4o correctly classify different sets of examples.**

5 Model Supervision

We introduce an end-to-end fine-tuned LLaVA (Language-Vision Alignment) model that rivals much larger models in performance. It offers edit evaluation abilities equivalent to SoTA models while significantly reducing computational costs, providing an efficient solution for AI-generated image edit evaluation.

We trained the model using the MagicBrush dataset which consists of 8,808 edits. A balanced set of 5,422 edits was used for artifact detection, while the full set was used for edit accuracy and caption generation. Two augmentation methods described below, produced 31,059 training instances per task. Further details are provided in Appendix A.9.

5.1 Negative Edit Augmentation

This method generates negative edits by selecting a deceptive target object and producing corresponding metadata, including instructions and difference captions. In Figure 7, a similarly sized scene object (an umbrella) was chosen as the deceptive target, and new metadata was generated using GPT-3.5 with few-shot prompting. For Add and Replace edits, the deceptive object is a visually similar absent object, like a cactus instead of a potted plant. For Change Attribute edits, attributes are modified, like altering a coat's color from blue to red.

5.2 Reverse Edit Augmentation

This augmentation focuses on reversing the edit using few-shot prompts with GPT-3.5. Add edits are changed to Remove edits, Replace edits involve

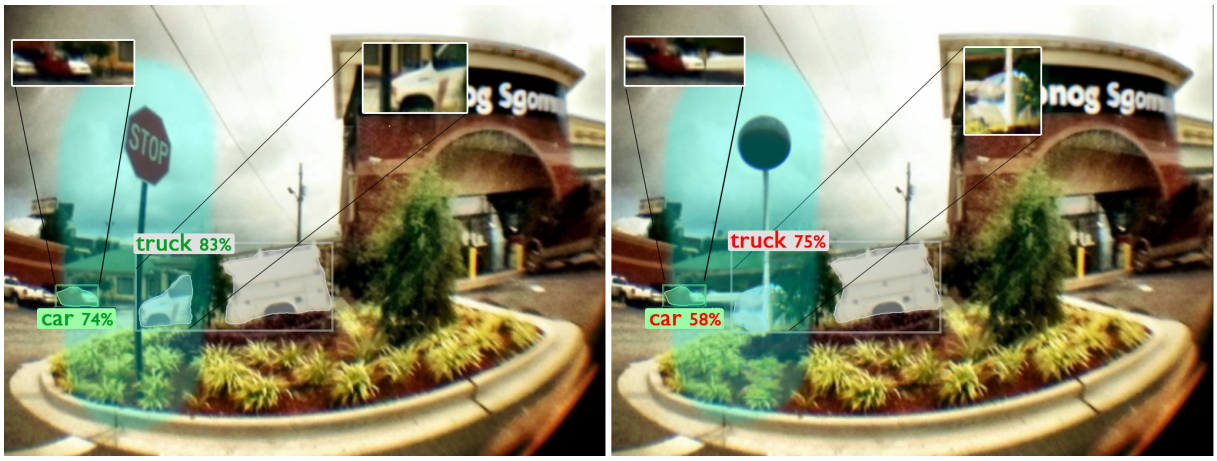

Figure 5: The first method for detecting artifacts using the Detic model for the edit "turn the stop sign to a lollipop". Comparing Detic probabilities for objects intersecting the turquoise in-painting mask between the pre-edit (left) and post-edit (right) images reveals two artifacts, the truck and small car, whose probability drops exceeds our threshold.

switching the source and target objects, and Change Attribute edits reverse the attribute modification. For example, in Figure 7, the edit "Remove one potted plant" is reversed to "Add one potted plant." Applied on top of the negative augmentation, this process expands the dataset fourfold, providing comprehensive training data for our model.

5.3 Supervision Results

Our model demonstrates competitive performance against SoTA VLMs. As shown in Table 3, it outperforms Gemini, GPT-4, and GPT-4-turbo in artifact detection, with only Gemini-1.5 (58.5%) and GPT-4o (65.7%) performing better. For Edit Accuracy, it achieves 67.2%, surpassing Gemini (49.9%) and GPT-4 Turbo. It also maintain competitive performance in the Difference Caption Accuracy (54.5%), surpassing Gemini model (53.9%). These results validate our augmentation methods and highlight the value of our training data.

6 Related Work

Recent advances in text-guided image editing enable modifications via natural language (Sheynin et al., 2023; He et al., 2024; Wu et al., 2023a; Cui et al., 2023), with some models supporting multi-turn refinement. Others use spatial masks for precise, localized edits (Avrahami et al., 2022; Nichol et al., 2022; Wang et al., 2023), which offer better control than text-only methods (Wang et al., 2023; Zhang et al., 2024a).

Edit quality is often measured using pixel-level similarity (L1/L2 norms) and CLIP-based cosine similarity (Radford et al., 2021). However, these metrics poorly align with human judgment (Basu et al., 2023), offering only quantitative scores without qualitative insights.

Image editing benchmarks like EditBench (Wang et al., 2023) and EditVal (Basu et al., 2023) assess editing models through automatic and human evaluations, focusing on instruction adherence and object or scene preservation. In contrast, our work evaluates models as edit inspectors on overlooked edit aspects such as scene integration, pixel-level issues, and artifact detection. We also introduce the category "Accurate, But Unexpected" to capture technically correct edits that deviate from user expectations and collect textual feedback and detailed difference captions to provide deeper insights into edit quality.

7 Conclusion and Future Work

In this work, we introduce EditInspector, a public benchmark for evaluating text-guided image edits using a comprehensive annotation framework. We also propose a zero-shot pipeline for instruction-grounded difference captions, a novel artifact detection method leveraging segmentation probabilities, and two augmentation techniques to generate synthetic training data for edit verification models. Future work can refine difference caption generation and explore new approaches to address existing model limitations.

We hope our benchmark and proposed methods for artifact detection, captioning, and augmentation drive advancements in edit evaluation and inspire further research in this domain.

8 Limitations

Our benchmark is based exclusively on the MagicBrush dataset for evaluating edits, which, while covering diverse scenarios, is limited to natural images and mask-guided edits. Recent studies have shown promising results with free-text methods (Sheynin et al., 2023) and growing interest in editing of synthetic images. Additionally, the distribution of edit types in the test set reflects the natural distribution of human edits from the MagicBrush dataset, as determined by a human study. While this mirrors real-world editing trends, it may not equally represent all edit types. These limitations highlight distinct research directions that could be explored independently of our current work.

References

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

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

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. *Preprint*, arXiv:2211.09800.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance. *Preprint*, arXiv:2210.11427.

Xing Cui, Zekun Li, Peipei Li, Yibo Hu, Hailin Shi, and Zhaofeng He. 2023. Chatedit: Towards multi-turn interactive facial image editing via dialogue. *Preprint*, arXiv:2303.11108.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Yingqing He, Zhaoyang Liu, Jingye Chen, Zeyue Tian, Hongyu Liu, Xiaowei Chi, Runtao Liu, Ruibin Yuan, Yazhou Xing, Wenhai Wang, Jifeng Dai, Yong Zhang, Wei Xue, Qifeng Liu, Yike Guo, and Qifeng Chen. 2024. Llms meet multimodal generation and editing: A survey. *Preprint*, arXiv:2405.19334.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. *Preprint*, arXiv:2210.09276.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *Preprint*, arXiv:2112.10741.

OpenAI. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. 2023. Emu edit: Precise image editing via recognition and generation tasks. *Preprint*, arXiv:2311.10089.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. *Preprint*, arXiv:2212.06909.

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

Tsung-Han Wu, Long Lian, Joseph E. Gonzalez, Boyi Li, and Trevor Darrell. 2023b. Self-correcting llm-controlled diffusion models. *Preprint*, arXiv:2311.16090.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2024a. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Preprint*, arXiv:2306.10012.

Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, Caiming Xiong, and Ran Xu. 2024b. Hive: Harnessing human feedback for instructional visual editing. *Preprint*, arXiv:2303.09618.

Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris Metaxas, and Jian Ren. 2022. Sine: Single image editing with text-to-image diffusion models. *Preprint*, arXiv:2212.04489.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting twenty-thousand classes using image-level supervision. In *ECCV*.

A Appendix

A.1 Common Caption Comparison Metrics

Common metrics for comparing image captions, such as BLEU, METEOR, and ROUGE, rely on N-gram overlaps between generated and reference texts. However, they fall short of our core requirement to ensure accurate alignment between the edited objects and actions described in the captions. As shown in Table 4, while these metrics suggest that GPT-4 generates captions most similar to the ground truth, in practice, it is the least accurate model, exhibiting the highest hallucination rate and the largest number of average changes detected. Below, we provide a brief explanation of these metrics, followed by several scenarios illustrating their limitations in effectively evaluating difference captions.

- BLEU: Computes the number of matches in unigrams, bigrams, trigrams, and 4-grams between generated and reference text. Includes a brevity penalty to discourage shorter outputs.
- ROUGE: ROUGE-1 calculates the F1 score for unigrams. ROUGE-2 calculates the F1 score for bigrams.
- METEOR: Incorporates features such as stemming, synonym matching, and paraphrase recognition. Computes the unigram F1 score.
- CIDEr: Measures the similarity between generated and reference captions using TF-IDF weighted n-grams (unigrams to 4-grams). Emphasizes consensus between generated captions and multiple human references while penalizing overuse of common n-grams.

Although these metrics are widely used in image captioning, they have severe limitations when evaluating difference captions for image edits.

Miss Weighting the Edited Objects and Actions. These metrics struggle to differentiate between critical objects and less significant words in the context of difference captions. For instance, consider the ground truth caption: "The main difference between the two images is the first image has a blue vase and the second image a brown vase." If the generated caption states, "The main difference between the two images is the first image has a squirrel and the second image does not," linguistic metrics might still assign relatively high scores (e.g., BLEU: 0.68, ROUGE-1 Recall: 0.81, METEOR: 0.78) due to superficial word overlaps. However, these scores fail to reflect the semantic misalignment between the captions. In contrast, our proposed metric assigns a score of 0, accurately reflecting the discrepancy in the identified edited object and action.

Accounting for Unchanged Objects, Varying Length, and Stylistic Differences. Conventional metrics often penalize captions that include mentions of unchanged objects, vary in length, or differ stylistically, even when accurately describing the detected changes. For instance, consider the generated caption: "The difference between the two images is that the first image has a blue vase. The second image has a blue vase and a squirrel next to it." Our metric would assign this caption a perfect score of 1, as it correctly identifies the key difference (the addition of the squirrel) in alignment with the ground truth caption: "A brown squirrel was added to the image." In contrast, linguistic metrics would score close to 0 due to the inclusion of details about the unchanged "blue vase" and penalties for variations in length and phrasing. This demonstrates the robustness of our metric in handling linguistic variability while focusing on the accuracy of detected changes.

Capturing the Order of Edits. The above mentioned metrics overlook the importance of edit sequence order. For instance, consider the ground truth captions: "In the first image, the tree was removed, and a new flowerbed was added" and the generated caption "In the first image, the flowerbed was removed, and a new tree was added." Although both captions involve the same objects (tree and flowerbed) and actions (added and removed), the sequence of edits conveys entirely different meanings. The n-gram based metrics would assign high scores to these captions because they mention the same words (objects and actions), regardless of their order, failing to penalize semantic misalignment. In contrast, our metric explicitly evaluates the edit sequence order, ensuring that generated captions accurately reflect the correct sequence of changes.

	Gemini-1.5	GPT-4	GPT-4o	GPT-4 Turbo
Differences Caption Generation				
Main Difference	31%	27%	**39.0%**	24%
MP	8%	8%	**12%**	8%
HR	67%	78%	**60%**	75%
METEOR	0.11	**0.22**	0.19	0.19
ROUGE-1	0.15	**0.36**	0.29	0.30
ROUGE-2	0.04	**0.09**	0.08	0.07
BLEU	0.01	0.02	**0.03**	0.02

Table 4: Comparison of models on the Difference Caption Generation task. GPT-4 achieves the best results on METEOR, ROUGE-1, and ROUGE-2 metrics, while GPT-4o ranks highest in BLEU.

A.2 Mathematical Explanation of Metrics

We evaluate model performance on **all differences captions** using two metrics: **Model Precision (MP)** and **Hallucination Rate (HR)**. These are computed based on Difference Triplets (DTs), defined as:

$$\text{DT} = (\text{source object}, \text{target object}, \text{action type}),$$

where *source object* is the original object affected by the edit, *target object* is the resulted object of the edit, and *action type* is the type of edit (e.g., "add," "remove," "replace"). **Model Precision (MP):** Measures the percentage of human-annotated DTs ($\mathcal{H}$) matched by model-detected DTs ($\mathcal{M}$):

$$\text{MP} = \frac{|\mathcal{H} \cap \mathcal{M}|}{|\mathcal{H}|} \times 100,$$

where $|\mathcal{H} \cap \mathcal{M}|$ is the number of matched DTs, and $|\mathcal{H}|$ is the total human-annotated DTs. **Hallucination Rate (HR):** Measures the percentage of model-detected DTs ($\mathcal{M}$) not matching any human-annotated DTs ($\mathcal{H}$):

$$\text{HR} = \frac{|\mathcal{M} \setminus \mathcal{H}|}{|\mathcal{M}|} \times 100,$$

where $|\mathcal{M} \setminus \mathcal{H}|$ is the number of hallucinated DTs, and $|\mathcal{M}|$ is the total model-detected DTs. **Soft Metrics:** MP_{soft} and HR_{soft} allow matches when source and target objects in DTs are reversed:

$$\text{MP}_{\text{soft}} = \frac{|\mathcal{H}_{\text{soft}} \cap \mathcal{M}|}{|\mathcal{H}|} \times 100,$$

$$\text{HR}_{\text{soft}} = \frac{|\mathcal{M} \setminus \mathcal{H}_{\text{soft}}|}{|\mathcal{M}|} \times 100.$$

Matching Criteria: A DT match requires identical *action type* and similar *source/target objects* (assessed by GPT-4). Relaxed matching ($\mathcal{H}_{\text{soft}}$) accounts for reversed source and target objects.

A.3 Metrics Example

We calculate the MP and HR metrics using Figure 1 GPT-4o and the human-annotated difference caption. The ground truth lists the following human-annotated differences ($\mathcal{H}$):

$$\begin{aligned}
&(\text{carpet floor}, \text{wooden floor}, \text{Replace}),\\
&(\text{None}, \text{door}, \text{Add}),\\
&(\text{fridge bottom}, \text{extended fridge bottom}, \text{Change}),\\
&(\text{yellow box}, \text{extended yellow box}, \text{Change}),\\
&(\text{yellow box text}, \text{None}, \text{Remove}),\\
&(\text{text}, \text{image}, \text{Replace})
\end{aligned}$$

GPT-4o detects only one difference:

$$\mathcal{M} = \{(\text{carpet floor}, \text{wooden floor}, \text{Replace})\}.$$

Model Precision (MP): Model Precision (MP) measures the percentage of human-annotated DTs ($\mathcal{H}$) matched by model-detected DTs ($\mathcal{M}$):

$$\text{MP} = \frac{|\mathcal{H} \cap \mathcal{M}|}{|\mathcal{H}|} \times 100.$$

The only match between $\mathcal{H}$ and $\mathcal{M}$ is:

$$(\text{carpet floor}, \text{wooden floor}, \text{Replace})$$

Therefore:

$$|\mathcal{H} \cap \mathcal{M}| = 1, \quad |\mathcal{H}| = 6,$$

$$\text{MP} = \frac{1}{6} \times 100 \approx 16.67\%.$$

Hallucination Rate (HR): Hallucination Rate (HR) measures the percentage of model-detected DTs ($\mathcal{M}$) that do not match any human-annotated DTs ($\mathcal{H}$):

$$\text{HR} = \frac{|\mathcal{M} \setminus \mathcal{H}|}{|\mathcal{M}|} \times 100.$$

Here, all model-detected DTs match human-annotated DTs, so:

$$\mathcal{M} \setminus \mathcal{H} = \emptyset, \quad |\mathcal{M}| = 1,$$

$$\text{HR} = \frac{0}{1} \times 100 = 0\%.$$

$$\text{MP} = 16.67\%, \quad \text{HR} = 0\%.$$

A.4 Additional Annotation Information

Each image edit was annotated by three annotators, with annotations conducted in batches of 27-54 edits. Annotators were paid at a rate of $0.70 per sample, resulting in an average hourly wage of $18.

To ensure the quality of annotations, we implemented a qualification test to select quality annotators. We provided detailed instructions, including decision trees that visually guide the answering process. These decision trees were accessible via the user interface ("tree icon"), allowing annotators to follow the guidelines while annotating image edits.

Additionally, a settings window was available, enabling annotators to customize the UI, including font size, width, and padding, to suit their personal preferences (See Appendix A.14).

A.5 Tasks Prompts

Model performance varied greatly with different prompts, suggesting that models may struggle to fully understand the task. We selected prompts that conveyed the user instructions and improved the overall performance.

- **Difference Caption Accuracy Task (Yes/No)**

  You are provided with before and after images of an image edit for the edit instruction "{}". Does the difference caption "{}" describe the difference between the two images (Answer only Yes/No)?

- **Visual Consistency Task (Yes/No)** You are provided with before and after images of an image edit for the edit instruction "{}". Is the edited object or its area (in remove/replace actions) consistent with the edit instruction and the image scene in terms of shape, size, brightness, shadows, texture, color, etc. (Answer only Yes/No)?

- **Is Accurate Task**

  You are provided with before and after images of an image edit for the edit instruction "{}". Was the edit instruction "{}" accurately executed and does it reflect the intended change (Answer only Yes/No)?

- **Artifacts Task**

  You are provided with before and after images of an image edit for the edit instruction "{}". Are there any artifacts or alterations in the image not intended to be affected by the edit "{}" (Answer only Yes/No)?

- **Technical Precision Task (Yes/No)**

  You are provided with before and after images of an image edit for the edit instruction "{}". Does the edited object or its area (in remove/replace actions) maintain the image resolution, exhibit blur, show any smoothness, etc. (Answer only Yes/No)?

- **Generate all differences caption** You are provided with before and after images of an image edit. Please describe all the differences between these two images. Focus only on the differences; do not include any irrelevant information. Ignore any style differences between the images, such as changes in artistic style, color grading, or filters.

- **Generate main differences caption** Please describe the main difference between the two images.

A.6 Textual Feedback

We compared the predicted feedback from the models with human annotations by using a zero-shot prompt with GPT-4o that determines whether two pieces of feedback share any common points (yielding a simple Yes or No). The models' feedback matched human feedback only in a very small percentage of cases. The contextual consistency feedback shared common points with human feedback in 7%-28% of cases, while technical precision feedback did so in 4%-46% of instances.

A.7 Categories of Feedback Issues

- **Shape/Proportion**: Captures distortions in the shape, size, or proportions of objects.
  - *Keywords*: shape, proportion, size, distorted, too big, too small
  - *Example*: "The bird has an odd shape and is also yellow."

- **Blur/Fuzziness**: Deals with visual issues related to blurred or unclear edges, lack of sharpness, and fuzziness.
  - *Keywords*: blurry, fuzzy, smudged, blurred edges, not clear
  - *Example*: "The cat's fur is smoothened and texture is changed."
- **Texture**: Focuses on objects with unrealistic or unnatural textures, often described as too smooth or grainy.
  - *Keywords*: texture, smooth, grainy, patchy, unnatural
  - *Example*: "The building texture is unnatural."
- **Lighting/Brightness**: Involves issues where shadows are inconsistent or missing, or where lighting is overexposed or underexposed.
  - *Keywords*: shadows, lighting, brightness, overexposed, underexposed
  - *Example*: "The white bright part on the pan gives it an unrealistic look."
- **Color**: Captures cases where colors are oversaturated, under-saturated, or do not align with the scene.
  - *Keywords*: color, too bright, saturated, unnatural color
  - *Example*: "The fox is bright and inconsistent with the rest of the image."
- **Unreal/Artificial Look**: Describes objects that appear cartoonish, toy-like, or overly artificial, failing to blend with the rest of the scene.
  - *Keywords*: cartoon, toy, artificial, fake, graphical
  - *Example*: "The helicopter's texture resembles a toy."
- **Placement**: Refers to objects that are misaligned or incorrectly oriented in the scene.
  - *Keywords*: placement, misaligned, incorrect angle, orientation
  - *Example*: "The curtain is hanging in the air instead of the bar."
- **Missing/Extra Objects**: Captures cases where objects are unexpectedly added or removed, causing inconsistencies.
  - *Keywords*: missing, removed, added, extra, inconsistent
  - *Example*: "The man's face was removed and replaced by a mask."
- **Edges**: Focuses on issues related to sharp, uneven, or poorly blended edges.
  - *Keywords*: edges, sharp, uneven, jagged
  - *Example*: "The edges of the pizza are not even."
- **Resolution**: Refers to cases where the visual clarity or quality of the image is degraded, often appearing pixelated or with visual noise.
  - *Keywords*: resolution, clarity, pixelated, low quality
  - *Example*: "The image of the bird looks pixelated and low in resolution."

A.8 Analysis Methodology

Our categorization process followed these steps:

1. **Examining the Workers' Feedback**: We reviewed detailed textual feedback from workers who evaluated the instruction-based edits. Each piece of feedback was carefully analyzed to identify recurring issues.
2. **Identifying Categories**: We identified common themes in the feedback and organized them into meaningful categories representing distinct visual and technical issues.
3. **Extracting Keywords for Categories**: For each category, we identified specific keywords and phrases that workers frequently used to describe the issues. These keywords were used to group similar feedback together.
4. **Generating Statistics**: We quantified the frequency of each category across the entire dataset to understand which types of issues were most prevalent. This analysis provided insights to guide future improvements in the edits.

A.9 Supervision Details

The model was fine-tuned for 1 epoch using AdamW with a 2×10^{-4} learning rate. Since it accepts a single image input, we concatenated the before-and-after images.

A.10 Model Versions

- **GPT Models:**
  - GPT-4o (2024-08-06)
  - GPT-4 Turbo (2024-04-09)
  - GPT-4 (0613)
- **Gemini Models:**
  - Gemini 1.5 Pro (001)
  - Gemini 1.0 Pro (001)

A.11 Additional Experiments

Figure 6 presents the precision and hallucination rates as a function of the number of objects in the edited images. There is a performance drop in all models as the number of objects in the images increases, highlighting a trend where more complex scenes contribute to higher hallucination rates and lower precision.

A.12 Augmentation methods

A.13 Licenses

All use of scientific artifacts is consistent with their intended use. This work focuses on evaluating existing models in the English language using images from the MagicBrush dataset and does not introduce new models, generate new images, or employ technologies that could pose ethical, societal, or safety risks. We collected anonymous human annotations using Amazon Mechanical Turk crowdsourcing platform. The images are used in accordance with the MagicBrush license, and the evaluation code and dataset are released under the CC-BY-4.0 license.

A.14 Annotation UI

A.15 Annotation Examples

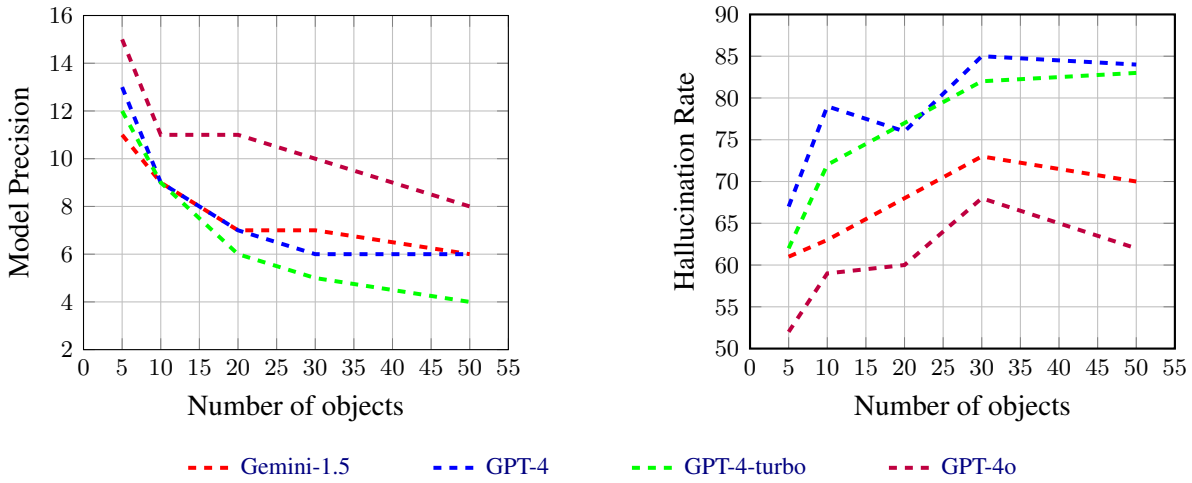

Figure 6: Comparison of model precision and hallucination rates as a function of the number of objects in the edited images. The performance of all models decreases as the number of objects in the images increases, highlighting a trend where more complex scenes contribute to higher hallucination rates and lower precision.

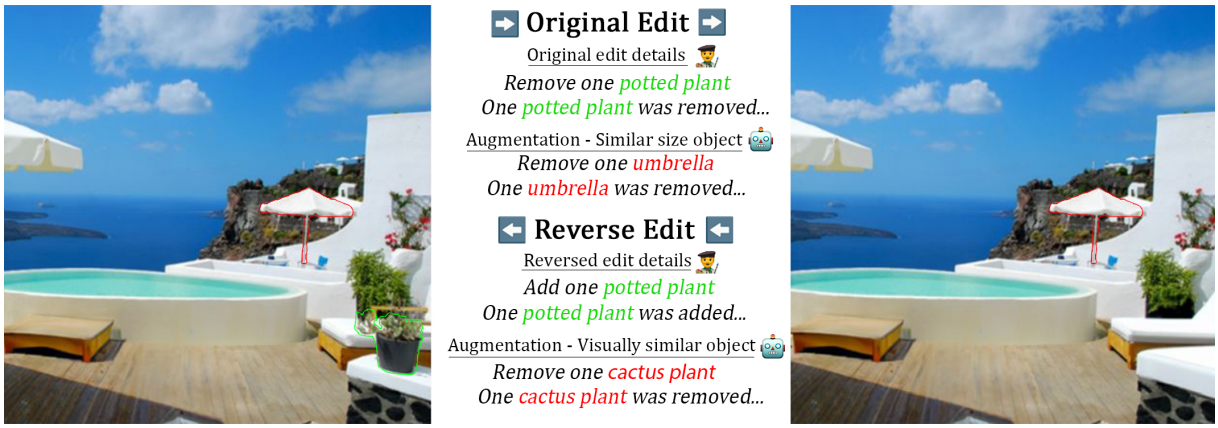

Figure 7: Illustration of our augmentation methods for a remove edit. The pre-edit image (left) shows a potted plant, while the post-edit image (right) depicts the scene with the plant removed. In the first augmentation method, the instruction and difference caption is modified by replacing the "potted plant" with an object of similar size (umbrella). In the second augmentation, we reverse the edit by switching the order of the images, changing the instruction and difference caption from "remove potted plant" to "add potted plant," and introducing a negative instruction for a visually similar object (e.g., cactus plant), which is absent in the post-edit image.

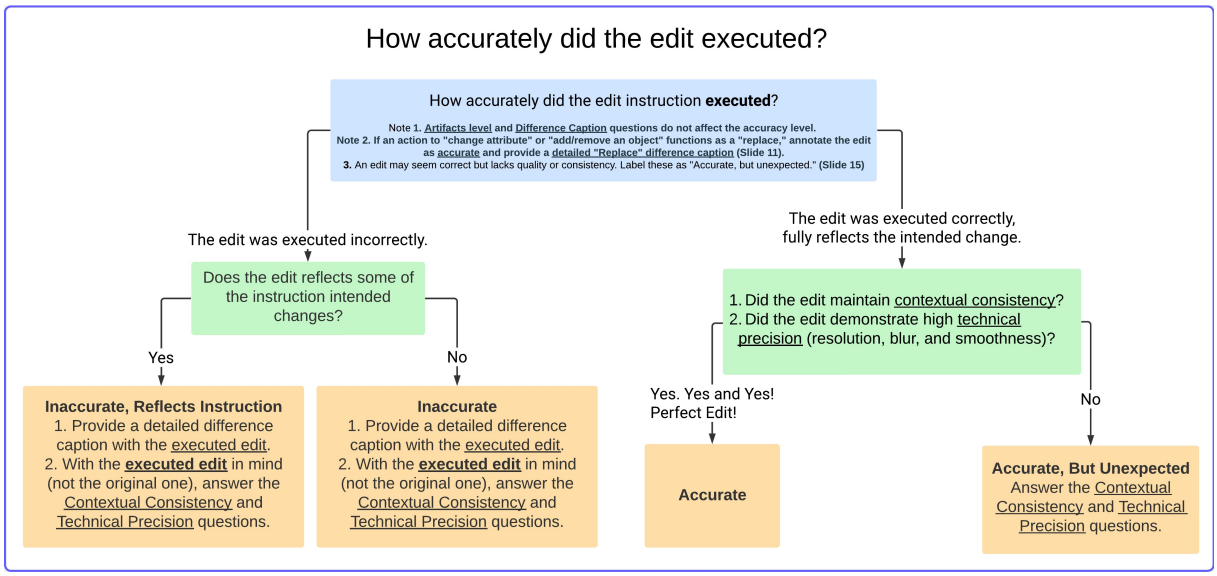

Figure 8: The accuracy scheme tree that was provided to annotators to guide the answering process.

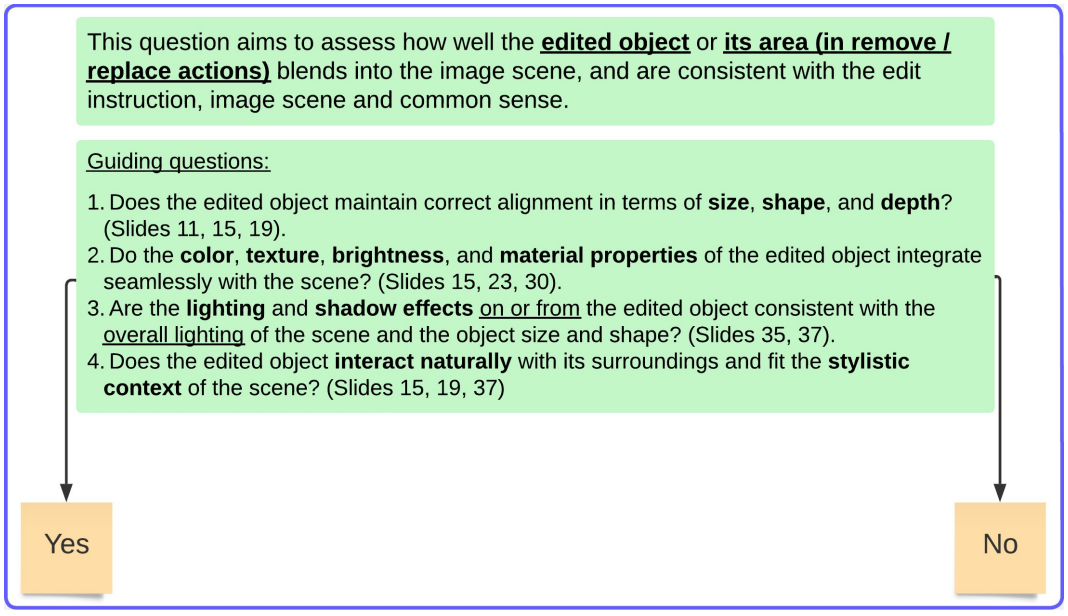

Figure 9: The contextual consistency scheme tree that was provided to annotators to guide the answering process.

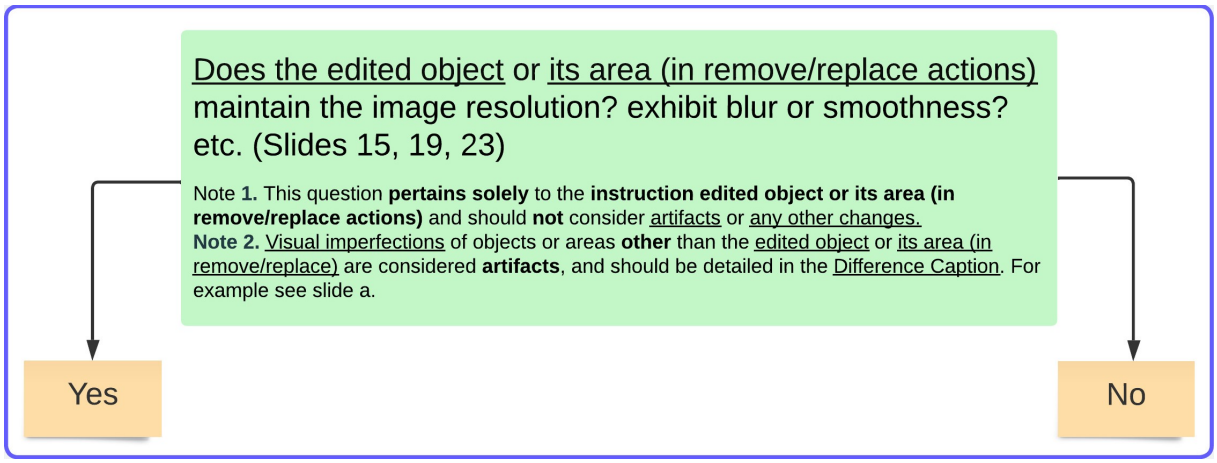

Figure 10: The technical precision scheme tree that was provided to annotators to guide the answering process.

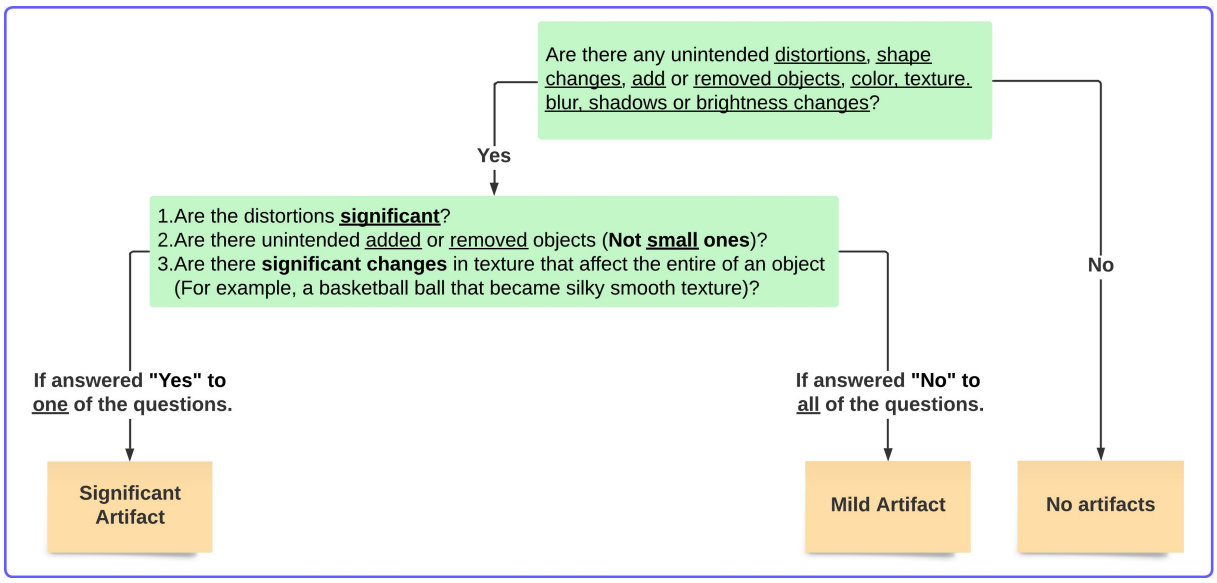

Figure 11: The artifacts scheme tree that was provided to annotators to guide the answering process.

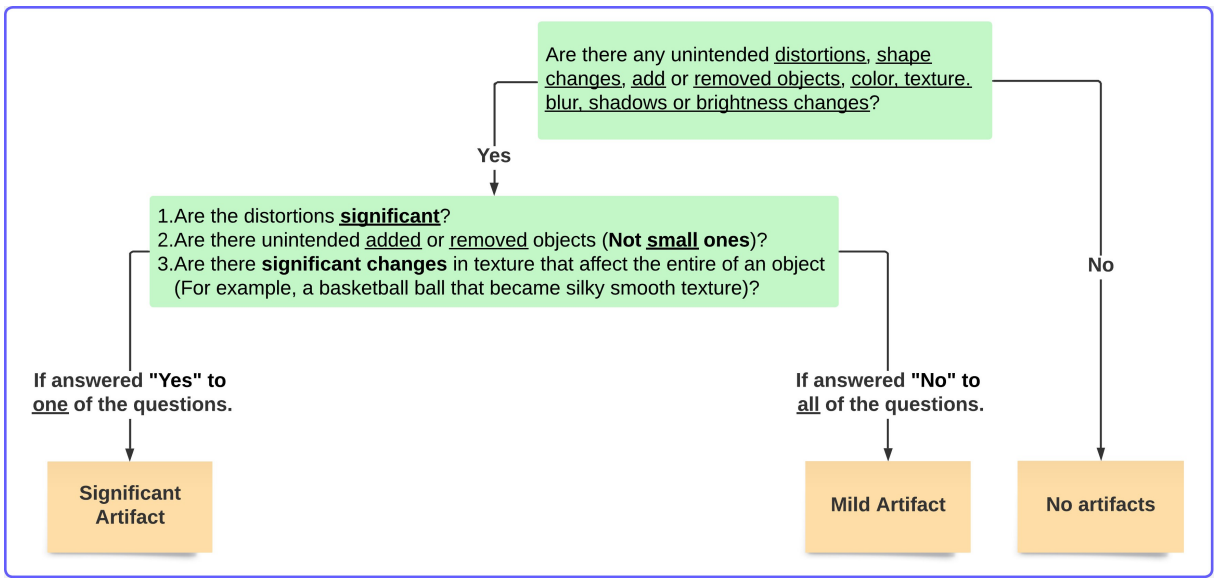

Figure 12: The difference caption instructions provided to annotators to guide the answering process.

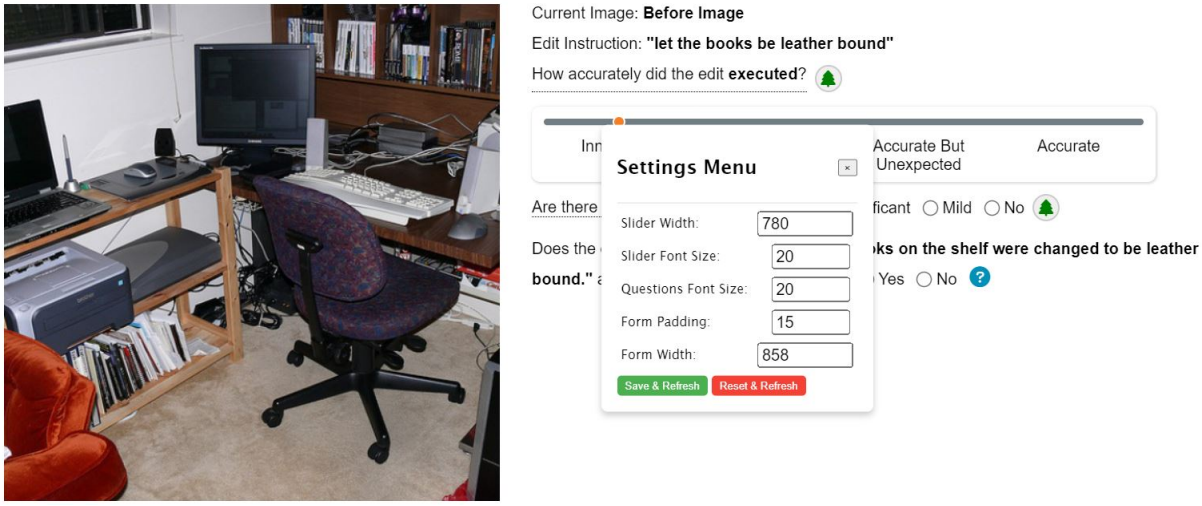

Figure 13: The setting menu for customizing the form font size, width etc.

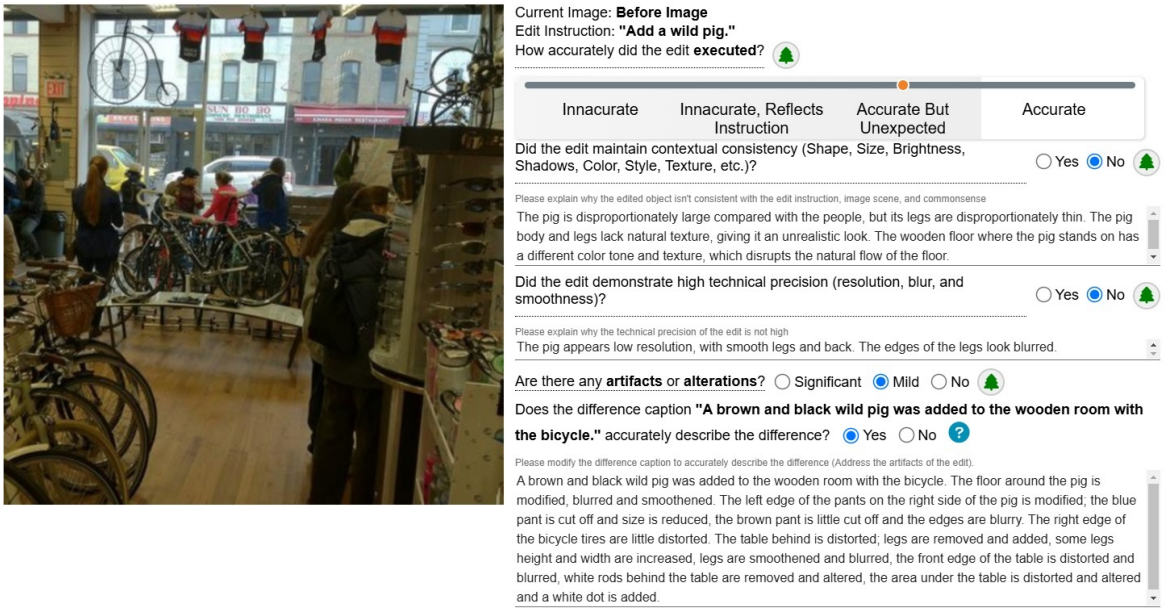

Figure 14: Example of image edit verification sample - before image (Add a wild pig).

Current Image: **After Image**
Edit Instruction: **"Add a wild pig."**
How accurately did the edit **executed**?

Innacurate | Innacurate, Reflects Instruction | Accurate But Unexpected | Accurate

Did the edit maintain contextual consistency (Shape, Size, Brightness, Shadows, Color, Style, Texture, etc.)? ○ Yes ◉ No

Please explain why the edited object isn't consistent with the edit instruction, image scene, and commonsense
The pig is disproportionately large compared with the people, but its legs are disproportionately thin. The pig body and legs lack natural texture, giving it an unrealistic look. The wooden floor where the pig stands on has a different color tone and texture, which disrupts the natural flow of the floor.

Did the edit demonstrate high technical precision (resolution, blur, and smoothness)? ○ Yes ◉ No

Please explain why the technical precision of the edit is not high
The pig appears low resolution, with smooth legs and back. The edges of the legs look blurred.

Are there any **artifacts** or **alterations**? ○ Significant ◉ Mild ○ No

Does the difference caption **"A brown and black wild pig was added to the wooden room with the bicycle."** accurately describe the difference? ◉ Yes ○ No

Please modify the difference caption to accurately describe the difference (Address the artifacts of the edit).
A brown and black wild pig was added to the wooden room with the bicycle. The floor around the pig is modified, blurred and smoothened. The left edge of the pants on the right side of the pig is modified; the blue pant is cut off and size is reduced, the brown pant is little cut off and the edges are blurry. The right edge of the bicycle tires are little distorted. The table behind is distorted; legs are removed and added, some legs height and width are increased, legs are smoothened and blurred, the front edge of the table is distorted and blurred, white rods behind the table are removed and altered, the area under the table is distorted and altered and a white dot is added.

Figure 15: Example of image edit verification sample - after image (Add a wild pig).

Current Image: **Before Image**
Edit Instruction: **"Can it be a cake on the plate?"**
How accurately did the edit **executed**?

Innacurate | Innacurate, Reflects Instruction | Accurate But Unexpected | Accurate

Did the edit maintain contextual consistency (Shape, Size, Brightness, Shadows, Color, Style, Texture, etc.)? ◉ Yes ○ No

Did the edit demonstrate high technical precision (resolution, blur, and smoothness)? ○ Yes ◉ No

Please explain why the technical precision of the edit is not high
The cake and affected area are much blurrier than the items replaced and doesn't match the focus of the sandwich that was originally on the plate. The edges of the cake are blurry.

Are there any **artifacts** or **alterations**? ○ Significant ◉ Mild ○ No

Does the difference caption **"The sandwich on the white plate was replaced with a piece of dark brown layered cake."** accurately describe the difference? ◉ Yes ○ No

Please modify the difference caption to accurately describe the difference (Address the artifacts of the edit).
The sandwich on the white plate was replaced with a piece of dark brown layered cake. The cake has gold frosting on top and between each layer. An upside down fork appeared under the person's thumb, and the white paper on the table widened and extended up to the plate.

Figure 16: Example of image edit verification sample - before image (Cake on the plate).

Current Image: **After Image**

Edit Instruction: **"Can it be a cake on the plate?"**

How accurately did the edit **executed**?

Innacurate | Innacurate, Reflects Instruction | Accurate But Unexpected | Accurate

Did the edit maintain contextual consistency (Shape, Size, Brightness, Shadows, Color, Style, Texture, etc.)? ○ Yes ○ No

Did the edit demonstrate high technical precision (resolution, blur, and smoothness)? ○ Yes ○ No

Please explain why the technical precision of the edit is not high

The cake and affected area are much blurrier than the items replaced and doesn't match the focus of the sandwich that was originally on the plate. The edges of the cake are blurry.

Are there any **artifacts** or **alterations**? ○ Significant ○ Mild ○ No

Does the difference caption **"The sandwich on the white plate was replaced with a piece of dark brown layered cake."** accurately describe the difference? ○ Yes ○ No

Please modify the difference caption to accurately describe the difference (Address the artifacts of the edit).

The sandwich on the white plate was replaced with a piece of dark brown layered cake. The cake has gold frosting on top and between each layer. An upside down fork appeared under the person's thumb, and the white paper on the table widened and extended up to the plate.

Figure 17: Example of image edit verification sample - after image (Cake on the plate).

Current Image: **Before Image**

Edit Instruction: **"delete the table"**

How accurately did the edit **executed**?

Innacurate | Innacurate, Reflects Instruction | Accurate But Unexpected | Accurate

Did the edit maintain contextual consistency (Shape, Size, Brightness, Shadows, Color, Style, Texture, etc.)? ○ Yes ○ No

Please explain why the edited object isn't consistent with the edit instruction, image scene, and commonsense

The shadow from one of the table legs is still there and somehow extended.

Did the edit demonstrate high technical precision (resolution, blur, and smoothness)? ○ Yes ○ No

Are there any **artifacts** or **alterations**? ○ Significant ○ Mild ○ No

Does the difference caption **"The white wicker coffee table with a glass top was removed from the image."** accurately describe the difference? ○ Yes ○ No

Please modify the difference caption to accurately describe the difference (Address the artifacts of the edit).

The white wicker coffee table with a glass top was removed from the image. The texture of the bottom of the curtain got altered. The texture and design pattern of the floor beneath the table got altered and now it has parallel lines instead of boxes. The texture of the couch behind it got altered along with the metal frame beside it.

Figure 18: Example of image edit verification sample - before image (Delete the table).

Current Image: **After Image**

Edit Instruction: **"delete the table"**

How accurately did the edit **executed**?

Innacurate | Innacurate, Reflects Instruction | Accurate But Unexpected | Accurate

Did the edit maintain contextual consistency (Shape, Size, Brightness, Shadows, Color, Style, Texture, etc.)? ○ Yes ◉ No

Please explain why the edited object isn't consistent with the edit instruction, image scene, and commonsense

The shadow from one of the table legs is still there and somehow extended.

Did the edit demonstrate high technical precision (resolution, blur, and smoothness)? ◉ Yes ○ No

Are there any **artifacts** or **alterations**? ○ Significant ◉ Mild ○ No

Does the difference caption **"The white wicker coffee table with a glass top was removed from the image."** accurately describe the difference? ◉ Yes ○ No

Please modify the difference caption to accurately describe the difference (Address the artifacts of the edit).

The white wicker coffee table with a glass top was removed from the image. The texture of the bottom of the curtain got altered. The texture and design pattern of the floor beneath the table got altered and now it has parallel lines instead of boxes. The texture of the couch behind it got altered along with the metal frame beside it.

Figure 19: Example of image edit verification sample - after image (Delete the table).

Current Image: **Before Image**

Edit Instruction: **"empty the table"**

How accurately did the edit **executed**?

Innacurate | Innacurate, Reflects Instruction | Accurate But Unexpected | Accurate

Are there any **artifacts** or **alterations**? ○ Significant ◉ Mild ○ No

Does the difference caption **"The table was emptied of its previous items: a round white plate, two glasses, and a white napkin."** accurately describe the difference?

◉ Yes ○ No

Please modify the difference caption to accurately describe the difference (Address the artifacts of the edit).

The table was emptied of its previous items: a round white plate, two glasses, and a white napkin. Black lines are added on the table top and the area around the items on the table is distorted. The left table edge is distorted. The black lines on the chairs behind the table are removed.

Figure 20: Example of image edit verification sample - before image (Empty the table).

Current Image: **After Image**

Edit Instruction: **"empty the table"**

How accurately did the edit **executed**?

Innacurate | Innacurate, Reflects Instruction | Accurate But Unexpected | Accurate

Are there any **artifacts** or **alterations**? ○ Significant ◉ Mild ○ No

Does the difference caption **"The table was emptied of its previous items: a round white plate, two glasses, and a white napkin."** accurately describe the difference?

◉ Yes ○ No

Please modify the difference caption to accurately describe the difference (Address the artifacts of the edit).

The table was emptied of its previous items: a round white plate, two glasses, and a white napkin. Black lines are added on the table top and the area around the items on the table is distorted. The left table edge is distorted. The black lines on the chairs behind the table are removed.

Figure 21: Example of image edit verification sample - after image (Empty the table).

Current Image: **Before Image**
Edit Instruction: **"let the man cut a pineapple"**
How accurately did the edit **executed**?

Innacurate | Innacurate, Reflects Instruction | Accurate But Unexpected | Accurate

Did the edit maintain contextual consistency (Shape, Size, Brightness, Shadows, Color, Style, Texture, etc.)? ○ Yes ◉ No

Please explain why the edited object isn't consistent with the edit instruction, image scene, and commonsense
The pineapple piece looks to be missing texture details.

Did the edit demonstrate high technical precision (resolution, blur, and smoothness)? ○ Yes ◉ No

Please explain why the technical precision of the edit is not high
The pineapple is blurred and of lower quality and resolution.

Are there any **artifacts** or **alterations**? ◉ Significant ○ Mild ○ No

Does the difference caption **"The beef roast that the man was slicing was replaced with a pineapple."** accurately describe the difference? ◉ Yes ○ No

Please modify the difference caption to accurately describe the difference (Address the artifacts of the edit).
The beef roast that the man was slicing was replaced with a pineapple. The two-pronged fork has been replaced with a knife, and a light reflection has been removed from the glass lid near the chopping board. The pattern of the stains on the chopboard changed. The handle of the chopboard got narrowed.

Figure 22: Example of image edit verification sample - before image (Cut a pineapple).

Current Image: **After Image**
Edit Instruction: **"let the man cut a pineapple"**
How accurately did the edit **executed**?

Innacurate | Innacurate, Reflects Instruction | Accurate But Unexpected | Accurate

Did the edit maintain contextual consistency (Shape, Size, Brightness, Shadows, Color, Style, Texture, etc.)? ○ Yes ◉ No

Please explain why the edited object isn't consistent with the edit instruction, image scene, and commonsense
The pineapple piece looks to be missing texture details.

Did the edit demonstrate high technical precision (resolution, blur, and smoothness)? ○ Yes ◉ No

Please explain why the technical precision of the edit is not high
The pineapple is blurred and of lower quality and resolution.

Are there any **artifacts** or **alterations**? ◉ Significant ○ Mild ○ No

Does the difference caption **"The beef roast that the man was slicing was replaced with a pineapple."** accurately describe the difference? ◉ Yes ○ No

Please modify the difference caption to accurately describe the difference (Address the artifacts of the edit).
The beef roast that the man was slicing was replaced with a pineapple. The two-pronged fork has been replaced with a knife, and a light reflection has been removed from the glass lid near the chopping board. The pattern of the stains on the chopboard changed. The handle of the chopboard got narrowed.

Figure 23: Example of image edit verification sample - after image (Cut a pineapple).