Human-Aligned MLLM Judges for Fine-Grained Image Editing Evaluation: A Benchmark, Framework, and Analysis

Anonymous ACL submission

Abstract

Evaluating image editing models remains challenging due to the coarse granularity and limited interpretability of traditional metrics, which often fail to capture aspects important to human perception and intent. Such metrics frequently reward visually plausible outputs while overlooking controllability, edit localization, and faithfulness to user instructions. In this work, we introduce a fine-grained Multimodal Large Language Model (MLLM)–as–a–Judge framework for image editing that decomposes common evaluation notions into twelve fine-grained interpretable factors spanning image preservation, edit quality, and instruction fidelity. Building on this formulation, we present a new human-validated benchmark that integrates human judgments, MLLM-based evaluations, model outputs, and traditional metrics across diverse image editing tasks. Through extensive human studies, we show that the proposed MLLM judges align closely with human evaluations at a fine granularity, supporting their use as reliable and scalable evaluators. We further demonstrate that traditional image editing metrics are often poor proxies for these factors, failing to distinguish over-edited or semantically imprecise outputs, whereas our judges provide more intuitive and informative assessments in both offline and online settings. Together, this work introduces a benchmark, a principled factorization, and empirical evidence positioning fine-grained MLLM judges as a practical foundation for studying, comparing, and improving image editing approaches.

1 Introduction

The rapid advancement of generative image editing models, with a highly impactful use-case being the ability to edit images via natural language prompts (Hertz et al., 2022; Kawar et al., 2023; Mokady et al., 2023; Zhang et al., 2023b; Ruiz et al., 2023; Shi et al., 2024; Couairon et al., 2022; Brooks et al., 2023), has introduced a critical challenge: how to reliably evaluate the quality of these edits. The traditional metrics relied on to assess image edit quality are fundamentally misaligned with human judgment in real-world applications. Addressing this problem requires a new evaluation system that captures semantic reasoning, contextual understanding, and alignment with human intent, rather than relying solely on pixel-level similarity.

Existing metrics such as Peak Signal-to-Noise Ratio (PSNR) (Jain, 1989), Structural Similarity Index Measure (SSIM) (Wang et al., 2004), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) usually focus on only one aspect of edit quality, such as low-level fidelity. Commonly used CLIP-based scores (Hessel et al., 2021) are also shown to be not reliable (Goel et al., 2022). They often fail to capture other important factors that humans care about, including context consistency, fine-grained semantics, and alignment with editing intent. As a result, these metrics may assign high scores to edits that appear unsatisfactory to human observers, or assign low scores to edits that appear satisfactory to human observers like the case presented in Figure 1. Refer to Appendix A for more detailed examples of traditional metrics failing.

Moreover, most traditional metrics depend on the availability of a ground-truth (GT) image, an idealized reference for comparison. In real-world scenarios, this type of reference rarely exists. Users typically provide only an instruction and an input image, expecting the model to generate a plausible output. As a result, conventional evaluation frameworks are poorly suited to these *online* (no-GT) settings, where performance must be assessed without comparison to a known correct image.

These limitations highlight the need for an evaluation framework that captures higher-level edit quality beyond pixel-level differences. Prior LLMs/MLLMs as judges approaches (Fu et al., 2024; Zheng et al., 2023; Hsu et al., 2023; Kim et al., 2024) show promise for subjective evaluation but

are not tailored to the unique demands of image editing. Existing MLLM-based judges for image editing approaches remain limited in that they either lack comprehensive multimodal reasoning for interpretable image edit assessment, being general on human judgments, or do not establish a unified scoring framework human evaluation and model-based judges (Basu et al., 2023; Pan et al., 2025; Yosef et al., 2025).

To address these gaps, we propose a fine-grained Multimodal LLM-as-a-Judge (MLLM Judge) framework that evaluates instruction-guided image edits along interpretable semantic factors, including contextual consistency, semantic correctness, image preservation, edit realism, and instruction fidelity. Our framework supports both offline (with ground truth) and online (no ground truth) settings, and explicitly enables a three-way comparison among human evaluation, MLLM-as-a-Judge, and traditional metrics. By jointly analyzing these perspectives, the MLLM Judge more closely reflects human assessment of image editing quality and delivers evaluations better aligned with human judgment than existing approaches.

Furthermore, we release a high-quality benchmark for others to leverage for fine-tuning, and other important use cases. Beyond its role as an evaluation framework, the benchmark and its 13 fine grained factors serve as a principled curriculum for improving MLLMs. By releasing fine grained, human annotated supervision, the benchmark provides the signal required to teach models why an edit is correct or incorrect, rather than merely what the final output should resemble. This structure naturally supports supervised fine tuning toward reasoning judges that produce chain of thought style critiques, grounding each score in explicit, factor level logic. In turn, the factorized design gives rise to factor aware RLHF, where dense, semantically meaningful rewards encourage a calibrated trade off between instruction fidelity and image preservation. Finally, the benchmark enables systematic hard negative mining and curriculum learning, allowing practitioners to surface concrete failure modes such as scale realism or spatial inconsistency and to train models along a controlled progression of increasingly complex editing behaviors.

Summary of Main Contributions. The key contributions of this work are as follows:

- **Benchmark.** We introduce a new image-editing benchmark that combines human evaluations over fine-grained factors (Table 1) with scores from models and traditional metrics, with representative use cases discussed in Appendix F.
- **Novel MLLM Judge Factors for Image Editing.** We propose an MLLM judge with 12 fine-grained factors spanning image preservation, edit quality, and instruction fidelity (Table 1), enabling diagnostic evaluation in both offline and online settings.
- **Human Alignment of Proposed Image Editing Judges**. We show strong alignment between human judgments and our MLLM judges across all factors and edit types, indicating reliable, human-consistent evaluation as shown in Table 2-3.
- **Traditional vs. Our MLLM Image Editing Judges.** We demonstrate that traditional metrics correlate with only a limited subset of our judge factors, while MLLM judges as detailed in Section 5 provide more intuitive and reliable assessments of edit quality.
- **Extensive Experiments & Findings.** We conduct comprehensive experiments and analyses, and release our code and data to support future research and practical adoption. The code and data are made available.[1]

2 Related Work

2.1 Image Editing Methods

Recent advances in text-guided image generation have enabled powerful multimodal models to synthesize images directly from natural language prompts (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Balaji et al., 2023; Ho et al., 2022), and many of these models have subsequently been adapted to support a wide range of image editing tasks (Hertz et al., 2022; Kawar et al., 2023; Mokady et al., 2023; Zhang et al., 2023b; Ruiz et al., 2023; Shi et al., 2024; Couairon et al., 2022; Brooks et al., 2023). As image editing increasingly relies on language-conditioned generative models, evaluating edit quality has become a fundamentally semantic and intent-driven problem. In parallel, prior work has explored LLMs/MLLMs as judges for subjective evaluation in domains such as UI quality (Luera et al., 2025), chart comprehension (Kim et al., 2025), text assessment (Kim et al., 2024; Fu et al., 2024), and even MLLM performance assessment (Zheng et al., 2023), demon-

[1]Our code and data: https://github.com/mllmasajudge-anonymous/MLLM-as-a-Judge

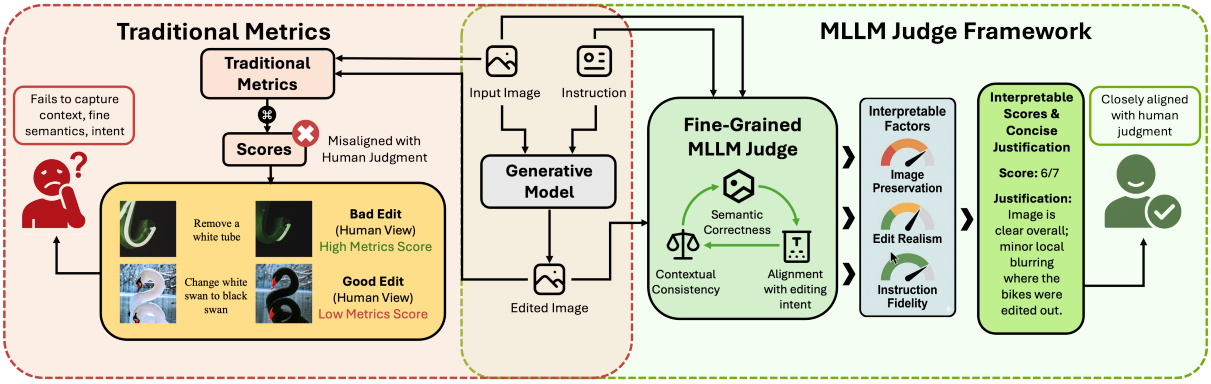

Figure 1: Motivation for fine-grained MLLM-based evaluation. The same image-editing example is assessed using two approaches: (left) traditional metrics, which collapse diverse editing behaviors into a single, potentially misleading score, and (right) our MLLM judge, which decomposes the edit into interpretable factors that explicitly explain why the edit succeeds or fails. This decomposition makes the motivation, methodology, and benefits of our approach immediately apparent.

Table 1: Our proposed taxonomy for image editing judges including higher-order categories of image preservation, edit quality, and instruction fidelity. We decompose these higher-order categories into finer-grained factors, including 3 factors for image preservation, 6 factors for edit quality, 3 factors for instruction fidelity, and lastly, an **Overall** factor, *totaling 13 factors*. For human evaluators, we include overall factor as a question, and for MLLM-as-a-Judge evaluators, the overall factor score is calculated from 12 other factors.

Category	Factor	Question
IMAGE PRESERVATION	Unchanged Regions	Did the parts of the image that were not supposed to be edited remain unchanged?
	Global Consistency	Has the overall appearance (style, layout, and color) been preserved?
	Identity Preservation	Do people, animals, or objects maintain their original identity and features after the edit?
EDIT QUALITY	Scale Realism	Is the scale of the edited object realistic compared to other objects in the image?
	Spatial Relationship	Has the spatial relationship between objects been maintained?
	Texture and Detail	Is the texture and detail in the edited region consistent with the surrounding areas?
	Image Quality	Does the edited image avoid noise, blurring, or unnatural distortions?
	Color and Lighting	Do the colors, shadows, and lighting of the edited region match the rest of the image?
	Seamlessness	Does the transition between edited and non-edited regions look natural?
INSTRUCTION FIDELITY	Alignment	Does the edited image align with the specific edits provided in the instructions?
	Completeness	Were all aspects of the instruction carried out fully?
	Plausibility	Does the result make sense in a real-world context?
Overall	Overall Edit Quality	Considering all factors, how good is the edit overall?

strating their ability to capture human preferences beyond static similarity metrics. More recent efforts have also applied MLLM-based judges to image editing evaluation (Fu et al., 2024; Zheng et al., 2023; Hsu et al., 2023; Kim et al., 2024). However, existing MLLM-based judges are not tailored to the unique failure modes of image editing. They typically provide coarse, holistic scores, instead of explicitly decomposing image editing quality into well-defined, separable dimensions. These limitations motivate a fine-grained, human-aligned MLLM-based evaluation framework specialized for image editing.

2.2 Traditional Metrics for Image Editing

Following prior work (Zhang et al., 2023a; Brooks et al., 2023; Bai et al., 2024), we adopt a set of traditional metrics grouped into three categories, as summarized in Table 4. Pixel-level fidelity metrics, including L1/L2 distance, PSNR (Jain, 1989), SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018), measure reconstruction accuracy by comparing edited images to ground truth. Content preservation metrics, such as Mask-SSIM and Mask-LPIPS, evaluate similarity without masked regions, while background consistency (Brooks et al., 2023) assesses whether unedited areas remain unchanged. Semantic alignment metrics capture higher-level relevance, with CLIP Score (Hessel et al., 2021) measuring text–image alignment in a reference-free manner and DINO similarity (Caron et al., 2021) comparing deep visual features against ground truth.

2.3 Factor Design Methods

Prior work consistently framed image editing evaluation as a multi-dimensional problem rather than a single metric. For example, EditVal (Dwivedi-

Table 2: **Human and our MLLM-as-a-Judge scores for all factors and across all edit types.** We report the average score over all image edit types in the last column and over all factors in the last row. When the difference between our judge score and human evaluation is closer than 0.5, its background is dark green. When the difference is closer than 1.0, its background is light green. The gray text is the standard deviation of the scores from which the average is computed.

	Factor		Image Edit Types						
			Add	Remove	Replace	Action	Counting	Relation	All Edits
Image Preserv.	Unchanged Regions	Human	5.172 (0.82)	5.731 (0.75)	4.972 (1.02)	4.352 (1.06)	3.393 (0.68)	4.992 (0.43)	4.769 (0.74)
		Our Judge	6.444 (0.50)	**5.824** (0.71)	6.222 (0.63)	5.826 (0.82)	4.400 (1.50)	**5.833** (0.90)	**5.758** (0.65)
	Global Consistency	Human	5.602 (0.70)	5.982 (0.61)	5.551 (0.90)	4.769 (0.87)	5.243 (1.13)	5.444 (0.39)	5.432 (0.37)
		Our Judge	**6.333** (0.47)	**5.971** (0.82)	**6.333** (0.47)	6.087 (0.65)	**4.800** (1.17)	**6.167** (0.69)	**5.948** (0.53)
	Identity Preservation	Human	5.613 (0.62)	5.913 (0.82)	5.625 (0.84)	4.871 (1.12)	4.227 (1.12)	5.714 (0.20)	5.327 (0.59)
		Our Judge	6.889 (0.31)	**6.118** (1.02)	**6.500** (0.96)	6.696 (0.46)	6.400 (0.49)	**6.500** (0.50)	6.517 (0.24)
Edit Quality	Scale Realism	Human	5.276 (0.95)	6.286 (0.54)	5.865 (0.80)	5.984 (0.61)	5.510 (0.68)	6.033 (0.57)	5.826 (0.34)
		Our Judge	6.444 (0.50)	**6.471** (0.74)	**6.556** (0.60)	**6.565** (0.92)	**6.400** (0.49)	**6.000** (1.41)	**6.406** (0.19)
	Spatial Relationship	Human	5.561 (0.79)	6.225 (0.50)	5.890 (0.63)	5.948 (0.74)	4.650 (1.19)	5.728 (0.63)	5.667 (0.50)
		Our Judge	**6.556** (0.68)	**6.382** (0.84)	**6.778** (0.63)	**6.043** (0.75)	**4.800** (1.17)	**6.667** (0.47)	**6.204** (0.67)
	Texture and Detail	Human	5.504 (0.56)	5.844 (0.61)	5.483 (0.80)	5.643 (0.84)	5.157 (0.90)	5.639 (0.59)	5.545 (0.21)
		Our Judge	**5.889** (0.31)	**5.676** (0.72)	**6.000** (0.47)	**6.000** (0.66)	**5.800** (0.40)	**5.667** (0.75)	**5.839** (0.14)
	Image Quality	Human	5.569 (0.54)	6.048 (0.66)	5.513 (0.71)	5.947 (0.71)	5.247 (0.53)	5.683 (0.75)	5.668 (0.27)
		Our Judge	**6.333** (0.47)	**6.059** (0.59)	6.556 (0.50)	**6.174** (0.48)	**6.200** (0.40)	**6.333** (0.75)	**6.276** (0.16)
	Color and Lighting	Human	5.515 (0.75)	5.855 (0.71)	5.442 (0.86)	5.549 (0.72)	5.403 (0.83)	5.553 (0.64)	5.553 (0.15)
		Our Judge	**6.111** (0.57)	**5.941** (0.87)	**6.278** (0.56)	**5.870** (0.80)	**6.200** (0.75)	**6.167** (0.69)	**6.094** (0.14)
	Seamlessness	Human	5.722 (0.64)	6.101 (0.58)	5.767 (0.74)	5.598 (0.83)	5.357 (1.00)	5.578 (0.73)	5.687 (0.23)
		Our Judge	**6.000** (0.47)	**5.706** (0.89)	**6.333** (0.58)	**5.913** (0.83)	**5.600** (0.80)	**5.667** (0.75)	**5.870** (0.25)
Instruct. Fidel.	Alignment	Human	5.556 (0.49)	5.927 (0.97)	5.681 (0.65)	5.666 (1.13)	3.437 (1.40)	5.178 (0.99)	5.241 (0.84)
		Our Judge	6.667 (0.94)	**6.471** (1.01)	**6.500** (0.90)	6.957 (0.20)	5.200 (2.23)	**6.167** (1.86)	6.327 (0.56)
	Completeness	Human	5.693 (0.72)	5.966 (1.17)	5.789 (0.71)	5.719 (1.14)	3.537 (1.74)	5.556 (0.66)	5.376 (0.83)
		Our Judge	6.778 (0.63)	**6.294** (1.30)	**6.389** (1.06)	6.870 (0.45)	5.200 (2.23)	**6.167** (1.86)	**6.283** (0.55)
	Plausibility	Human	5.209 (1.03)	6.023 (0.78)	5.743 (0.80)	5.692 (1.17)	4.917 (1.08)	5.586 (0.89)	5.528 (0.36)
		Our Judge	6.667 (0.47)	**6.529** (0.78)	6.889 (0.31)	6.826 (0.64)	6.800 (0.40)	**6.500** (0.76)	6.702 (0.15)
	Overall Average	Human	5.499 (0.17)	5.992 (0.15)	5.610 (0.24)	5.478 (0.50)	4.673 (0.78)	5.557 (0.26)	5.652 (0.47)
		Our Judge	**6.426** (0.30)	**6.120** (0.29)	**6.444** (0.23)	**6.319** (0.41)	**5.650** (0.74)	**6.153** (0.31)	**6.236** (0.40)

Table 3: Results comparing the overall scores from our MLLM judges to humans across the higher-order categories of image preservation, edit quality, and instruction fidelity.

Category		Image Edit Types						
		Add	Remove	Replace	Action	Counting	Relation	All Edits
Image Preserv.	Human	5.462 (0.71)	5.875 (0.73)	5.383 (0.92)	4.664 (1.02)	4.288 (0.98)	5.383 (0.34)	5.176 (0.57)
	Our Judge	6.555 (0.43)	**5.971** (0.85)	**6.352** (0.69)	6.203 (0.64)	**5.200** (1.05)	**6.167** (0.70)	**6.074** (0.47)
Edit Quality	Human	5.524 (0.70)	6.060 (0.60)	5.660 (0.76)	5.778 (0.74)	5.221 (0.85)	5.702 (0.65)	5.658 (0.28)
	Our Judge	**6.222** (0.50)	**6.039** (0.77)	**6.417** (0.56)	**6.094** (0.74)	**5.833** (0.67)	**6.084** (0.80)	**6.115** (0.26)
Instruct. Fidel.	Human	5.486 (0.75)	5.972 (0.97)	5.738 (0.72)	5.692 (1.15)	3.964 (1.41)	5.440 (0.85)	5.382 (0.68)
	Our Judge	6.704 (0.68)	**6.431** (1.03)	**6.593** (0.76)	6.884 (0.43)	5.733 (1.62)	**6.278** (1.49)	6.437 (0.42)
Overall Average	Human	5.499 (0.17)	5.992 (0.15)	5.610 (0.24)	5.478 (0.50)	4.673 (0.78)	5.557 (0.26)	5.652 (0.47)
	Our Judge	**6.426** (0.30)	**6.120** (0.29)	**6.444** (0.23)	**6.319** (0.41)	**5.650** (0.74)	**6.153** (0.31)	**6.236** (0.40)

Yu et al., 2022) assessed performance across edit quality, preservation of irrelevant content, and overall image consistency, aligning closely with our taxonomy of Instruction Fidelity, Image Preservation, and Edit Quality. MagicBrush (Zhang et al., 2023a) further emphasized human-centered evaluation, highlighting instruction consistency and overall visual quality as key factors. More recently, Pico-Banana-400K (Qian et al., 2025) introduced a structured automated pipeline that incorporated criteria such as instruction compliance, preservation balance, seamlessness, and technical quality.

Together, these benchmarks reflected a clear consensus: reliable evaluation must jointly consider faithfulness to instructions, preservation of original content, and visual realism. Building on this foundation, our framework unified these dimensions into a more fine-grained and interpretable factor space, enabling more precise error diagnosis and improving the reliability and consistency of automated judges. For additional related works, see Appendix B.

3 Fine-Grained MLLM Judges

In this section, we introduce our novel framework of fine-grained MLLM judges for evaluating instruction-guided image editing. This framework formalizes human-aligned evaluation by decomposing edit quality into interpretable semantic dimensions. First, we formulate the problem and motivate the need for a fine-grained approach (Section 3.1), and then we detail our proposed set of MLLM judge factors (Section 3.2).

3.1 Problem Formulation

We formalize the image-editing evaluation problem as deriving a scalar score $S(I_o, I_e, T)$ consisting of an *original image* I_o, a corresponding *edited image* I_e, and a natural language *instruction* T. The score $S(I_o, I_e, T)$ must reflect a balance between two competing objectives: adherence to the instruction and preservation of the original content.

While this objective can be captured by a single, general evaluation function, we argue that such formulations are insufficient for diagnosing image-editing behavior. Instead, our work introduces a **fine-grained MLLM judge**, defined as an evaluation function that decomposes S into interpretable sub-scores, where each sub-score s_i quantifies a distinct, separable aspect of edit quality. This decomposition is meaningful because image-editing failures are often nuanced: a high-fidelity edit can fail on subtle issues like lighting consistency, and traditional single-score metrics cannot distinguish these failure modes. By leveraging an MLLM's reasoning capabilities, we can automatically produce these sub-scores, providing both an overall quality assessment and a breakdown of *when* and *how* an editing approach succeeds or fails. Our formulation supports both the **online setting** (evaluating I_e against I_o and T) and the **offline setting** (where a ground truth I_g can be used for error analysis).

3.2 Our MLLM Judge Factors

To achieve fine-grained, interpretable assessment, we introduce a set of twelve key **judge factors** for image editing tasks, which are organized into three fundamental, high-level categories that cover the full spectrum of image editing success: *Image Preservation*, *Edit Quality*, and *Instruction Fidelity* as provided in Table 1. Each factor is scored using a 7-point Likert scale. An overview of these factors with examples is provided in Figure 2.

3.2.1 Image Preservation

This category evaluates the crucial requirement that a model must preserve regions of the image that are not targeted by the instruction. Failures in this category typically indicate over-editing, where unedited content is altered in ways that undermine the user's intent.

Unchanged Regions: Evaluates whether image regions not specified or implied by the instruction remain unchanged after editing.

Global Consistency: Evaluates if the scene's background, style, composition, and color palette remain consistent with the original image outside of the edited area.

Identity Preservation: Ensures primary subjects not involved in the edit retain their original identity and recognizable features.

3.2.2 Edit Quality

This category evaluates the visual realism and technical correctness of the edited content itself, independent of instruction compliance or preservation of unedited regions. This set is inspired by classic artistic and image analysis criteria concerning composition and visual coherence.

Scale Realism: Assesses whether the edited object or region has a realistic size and proportion relative to the scene context and depth.

Spatial Relationship: Evaluates whether edited elements maintain correct spatial relationships and perspective with surrounding objects.

Texture and Detail: Checks whether textures and fine details in the edited region are realistic and consistent with the surrounding image.

Image Quality: Determines whether the edited image avoids visible artifacts such as noise, blur, or unnatural distortions.

Color and Lighting: Assesses whether the colors, shadows, and lighting of the edited region are consistent with the scene's illumination.

Seamlessness: Evaluates whether transitions between edited and non-edited regions are smooth and visually natural.

3.2.3 Instruction Fidelity

This category evaluates whether the edited image correctly follows the given textual instruction and reflects the user's intended semantic content, moving beyond simple pixel comparisons to capture linguistic and contextual understanding.

Factor	Bad Edit	Score	Justification	Good Edit	Score	Justification
Unchanged Regions		1	Desert landscape replaced by an underwater scene; large unrelated background areas were changed.		7	Background landscape, road, and distant buttes remain identical between images without visible alteration.
Global Consistency		1	Photorealistic photo turned into vivid painting; style and color scheme drastically differ.		7	Overall color palette, contrast, and layout are preserved; scene appearance matches across both images.
Identity Preservation		1	Original boy replaced by faceless mannequin — hair, clothing, and facial features lost.		7	Boy's hair, clothing, skin markings, and body shape are unchanged and clearly preserved.
Factor	**Bad Edit**	**Score**	**Justification**	**Good Edit**	**Score**	**Justification**
Scale Realism		1	Hands are massively oversized compared to the boy and surrounding environment; scale is unrealistic.		7	Arm length and body proportions remain realistic and consistent with the surrounding environment.
Spatial Relationship		2	Boy stands on a different rock with altered distance and alignment to background mesas and paths.		7	Boy's position on the ridge and spatial relation to foreground vegetation are unchanged and consistent.
Texture and Detail		2	Cartoon lacks skin, fabric, and hair texture; flat colors contrast with detailed environment.		7	Shirt folds, skin texture, and ground details maintain consistent sharpness and fine detail.
Image Quality		2	Image shows severe grain and blur throughout the frame, degrading overall image quality.		7	No added noise, blurring, or compression artifacts; image remains sharp and clean.
Color and Lighting		2	Artificial neon hues and inconsistent shading conflict with the natural desert lighting and shadows.		7	Shadows and highlights on the arms and shirt match scene lighting and color temperature.
Seamlessness		2	Transitions are abrupt; thick black boundaries around arms create clearly visible seams against the background.		7	Transitions at the wrists and sleeves show no visible seams, blending naturally with surrounding pixels.
Factor	**Bad Edit**	**Score**	**Justification**	**Good Edit**	**Score**	**Justification**
Alignment		1	Instruction requested stretched-out hands, but the edited image shows the boy standing with hands down by sides.		7	Edited image displays the boy with his hands stretched outward as specified in the instruction.
Completeness		3	Instruction to have the boy stretch out his hands is only partially executed—left arm not extended.		7	Both arms are extended symmetrically and the requested pose change is fully implemented.
Plausibility		2	Floating arms with black voids are anatomically implausible and visually unrealistic.		7	The pose appears natural and physically plausible within the outdoor ridge setting.

Figure 2: Overview of the proposed factors used in our MLLM-as-a-Judge for image editing. Results are shown for each factor using both poorly edited images and those that were edited well (implementation in Fig. 8).

Alignment: Assesses whether the specific edit type and target described in the instruction are correctly realized in the edited image.

Completeness: Evaluates whether all components and constraints of the instruction are fully executed, rather than partially fulfilled.

Plausibility: Assesses whether the edited result is visually and physically plausible assuming a generally reasonable instruction, rather than judging the realism of the instruction itself.

3.3 Base Models and Implementation Details

Our fine-grained MLLM judges can be built upon any base MLLM. In this work, unless otherwise specified, we use GPT-5-mini as the base model, and all results leverage the general implementation provided in Appendix G (Figure 8). This is one of the simplest implementations of our approach, designed to be applicable across different evaluation settings and constitutes the default throughout the paper, unless otherwise mentioned. Nevertheless, we investigated other implementations of our fine-grained MLLM judges, and report results; please see Appendix G (Fig. 7-9), and for other base model results, see Appendix 5.3.

4 Methodology

This section describes our methodology for collecting our benchmark data and evaluation methodology of our fine-grained MLLM judges.

4.1 Benchmark Collection

To curate our benchmark, we selected 100 image editing tasks from the HumanEdit data. More specifically, we sampled uniformly at random 100 (original image, instruction) pairs, spanning six distinct edit types to ensure comprehensive coverage: Add (9 samples), Remove (34 samples), Replace (18 samples), Action (23 samples), Counting (10 samples), and Relation (6 samples). We evaluate our MLLM judge in an online setting, where edited images are generated rather than using pre-existing ground truth edits. For each pair, we used gpt-image-1 to generate an edited image based on the instruction, resulting in 100 (original, instruction, edited) triplets for evaluation. For more insight, refer to Appendix F. By proposing a set of fine-grained factors shown in Table 1 and grounding them in human annotations, we created a gold standard that can be used to evaluate if an MLLM judge actually thinks like a human at the finest granularity.

4.2 Participants and Procedure

Recruitment. We recruited 25 annotators representing diverse demographics, including undergraduate and graduate students, early-career professionals, and experienced practitioners. This diversity ensures our evaluation captures varied perspectives on image quality and editing success across different user backgrounds and expertise levels.

Sample Size and Coverage. Our study evaluates 100 image editing tasks, each comprising an original image, an editing instruction, and an edited image. For reliability analysis, each task is rated by five annotators. We recruited 25 participants, with each evaluating 20 randomly sampled images spanning six edit types.

Evaluation Structure. Every image editing task was judged using the 12 factors described in Section 2 and summarized in Table 1. Participants also provided an overall quality rating, yielding 13 scores per image (12 factors + 1 overall). Each participant provided 13 scores for each of their 20 assigned images, generating 260 scores per participant (13 scores × 20 images).

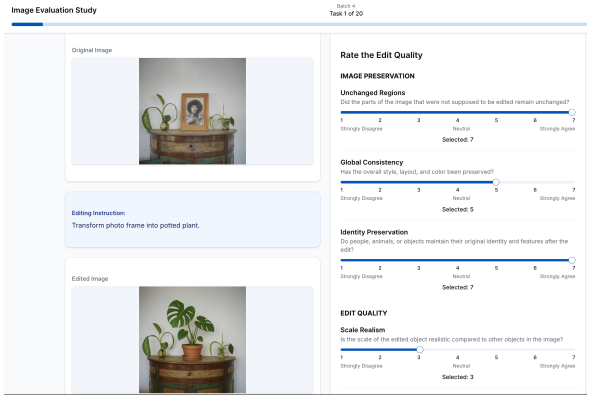

Figure 3: Human evaluation interface for image editing: evaluators view the original image, instruction, and edited result, then rate multiple dimensions using Likert scales.

4.3 Human Study Interface

Figure 3 shows our evaluation interface. This top-to-bottom arrangement follows natural reading flow and allows participants to easily compare before and after states while keeping the instruction visible. A progress indicator at the top displays "Task X of 20" to track evaluation progress.

4.4 Metrics

We compute several metrics to assess both human judgment characteristics and human-MLLM alignment. For human evaluation, we analyze mean and standard deviation of scores per factor. For human-MLLM alignment, we compute: (1) score prediction accuracy using Mean Squared Error (MSE) and Mean Absolute Error (MAE), (2) ranking correlation using Pearson, Spearman's ρ, and Kendall's τ, (3) agreement rates measuring percentage of cases where MLLM and human scores differ by $\leq$1 point. Further details for all the metrics are provided in Appendix D.3.

5 Experiments

We evaluate the effectiveness of MLLMs and our fine-grained judges for assessing image editing tasks by comparing their judgements to human judgements collected through our study (Section D). We compare the effectiveness of our approaches in capturing all the image editing factors proposed in Table 1.

5.1 Experimental Setup

For this work, we construct our benchmark based on the HumanEdit dataset (Bai et al., 2024), which provides high-quality human-annotated image editing data. The benchmark is used throughout all subsequent evaluations, with additional details on its construction, tasks, and usage provided in Appendix F.1. Our benchmark covers six edit types, namely add, replace, remove, counting, action, and relation. Each edit type targets a distinct class of image editing behaviors, and their detailed definitions are provided in Appendix F.3.

5.2 Results

We evaluate the effectiveness of our fine-grained MLLM judge by comparing its scores against human judgments across the 12 fine-grained factors in Table 1. Scores were averaged across annotators for each edited image.

In Table 2, we provide a detailed comparison between human judges and our fine-grained MLLM judges across the 12 proposed fine-grained image editing factors related to image preservation, edit quality, and instruction fidelity, stratified by *edit type*. The results in Table 2 demonstrate a high level of alignment between the proposed MLLM judges and human evaluators across diverse types of image edits. In particular, across nearly all factors and image edit types, our approach is shown to align strongly with human evaluators while providing more consistent scoring behavior, especially for complex edit types such as remove, replace, and relation. Furthermore, our fine-grained judges align closely with human evaluations across all three factor groups, with especially strong agreement on image preservation and edit quality. The judge maintains stable performance for instruction fidelity under challenging edit types, giving rise to

comparable or higher aggregate averages despite increased variance at the per edit level.

As shown in Table 3, the proposed MLLM judge achieves a high degree of alignment with human evaluators across image editing factor categories (image preservation, edit quality, and instruction fidelity); see Table 1 for the specific fine-grained factors under each category. Furthermore, our judge closely tracks human evaluations across all three factor groups, with especially strong agreement on image preservation and edit quality.

We now discuss several other findings we observed. More specifically, Table 2 reveals that the judge has strong performance on edit quality with almost all the blocks shaded green. This means that the judge captured many of the same visual cues and criteria that humans relied on when assessing edited images in this category. Moreover, the relatively small gray standard deviations accompanying the judge scores in these aligned regions indicated stable and confident judgments, often comparable to or even lower than those of human annotators. This was particularly evident in task types Add, Replace, and Counting, where agreement between human and judge evaluations was frequent and the judge exhibited low variance across all factors. The judge also perform well on image preservation, with global consistency task nearly all shaded green. Even some of the block have large difference for unchanged regions, the all edits column still show strong alignment. Even for more challenging instruction fidelity aspects, several edit types showed close agreement with human scores, suggesting that the judge reliably assessed not only perceptual quality but also semantic adherence to instructions. In addition, our results from Appendix Section I, J, and K showed that the benchmark method exhibited low inter-factor correlation, aligned well across human evaluators, and ensured the benchmark covering diverse quality levels.

5.3 Ablation Study

5.3.1 Varying LLM

When designing our MLLM judge, we evaluate two mainstream MLLMs, GPT-5-mini and Gemini-2.5-pro. Results for both models are reported in Table 12. Overall, GPT-5-mini shows consistently closer alignment with human evaluations than Gemini-2.5-pro across most factors, edit types, and the aggregated score. Nevertheless, Gemini-2.5-pro outperforms GPT-5-mini on specific factors and tasks, such as the Unchanged Regions factor in the Add task. More comprehensive results for all implementations are provided in Appendix H.

5.3.2 Varying Judge Instructions

We also investigated a variety of different implementations of our fine-grained MLLM judges.

- **Main (Figure 7):** This implementation only includes the most basic instructions.
- **Factor-level Rubrics (Figure 8):** This implementation includes fine-grained scoring rubrics for each judge factor.
- **Category wise, Example guided (Figure 9):** This implementation groups factors into three categories and uses category-specific prompts with detailed rubrics and examples.

As shown in Table 2, Table 5, and Table 6, the prompt illustrated in Figure 7 consistently achieves the best result. Additional result analysis are provided in Appendix H. Among all traditional metrics, CLIP and DINO Image performs the best bust still only present limited alignment with human evaluation result, showing the need for a more suitable evaluator. Full quantitative results and analysis are provided in Appendix C.

6 Conclusion

This work advances the evaluation of image editing approaches by moving beyond coarse and opaque metrics toward a fine grained, human aligned judging paradigm. By decomposing image editing quality into twelve interpretable factors spanning image preservation, edit quality, and instruction fidelity, we provide a principled lens for understanding when and why image edits succeed or fail. The accompanying human validated benchmark, which unifies human judgments, MLLM based evaluations, model outputs, and traditional metrics, enables systematic and reproducible analysis across a diverse set of image editing tasks. Our empirical findings demonstrate that MLLM judges align closely with human evaluations at a fine granularity, while commonly used metrics are often ineffective proxies for the aspects that matter most to users. Together, these contributions establish a practical foundation for diagnosing, comparing, and improving image editing approaches, and suggest a clear path toward more interpretable, reliable, and human aligned evaluation in both offline benchmarking and online development settings.

7 Limitations

While this work introduces a fine-grained, human-aligned MLLM-as-a-judge framework for image editing, several limitations remain. First, our benchmark and evaluations focus on a fixed set of image editing tasks and factor definitions, which, although representative, may not exhaustively capture all real-world editing scenarios or emerging edit types. Second, our fine-grained judges leverage state-of-the-art MLLMs, and their behavior naturally reflects the capabilities of these models at the time of evaluation. As MLLMs continue to improve, the proposed framework can directly benefit from improved reasoning and perception. Finally, while we observe strong alignment with human judgments, our study does not eliminate the need for humans in high-stakes or subjective applications, and further investigation is required to understand failure modes and edge cases.

7.1 Potential Risks

Over-reliance on human evaluator benchmarks introduces several potential risks. It may encode annotator subjectivity and cultural bias, encourage models to overfit benchmark-specific preferences, and mask inter-annotator disagreement through aggregated scores. Moreover, systems optimized for human benchmarks may avoid unconventional yet valid edits that deviate from annotator expectations, thereby discouraging creativity and limiting generalization to novel editing behaviors.

8 Ethical Considerations

The human evaluations were conducted with informed consent and appropriate quality controls. While MLLM-based judges may reflect biases in their underlying models or training data, we position our framework as a diagnostic and evaluative tool, and encourage careful use and continued auditing when applied in sensitive or high-stakes image editing settings.

References

Jinbin Bai, Wei Chow, Ling Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Shuicheng Yan. 2024. Humanedit: A high-quality human-rewarded dataset for instruction-based image editing. *arXiv preprint arXiv:2412.04280*.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. 2023. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

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. *arXiv preprint arXiv:2310.02426*.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

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

Jane Dwivedi-Yu, Timo Schick, Zhengbao Jiang, Maria Lomeli, Patrick Lewis, Gautier Izacard, Edouard Grave, Sebastian Riedel, and Fabio Petroni. 2022. Editeval: An instruction-based benchmark for text improvements. *arXiv preprint arXiv:2209.13331*.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. 2022. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528.

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. 2022. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33.

Ting-Yao Hsu, Chieh-Yang Huang, Ryan Rossi, Sungchul Kim, C Lee Giles, and Ting-Hao Huang. 2023. Gpt-4 as an effective zero-shot evaluator for scientific figure captions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5464–5474.

Anil K Jain. 1989. *Fundamentals of digital image processing*. Prentice-Hall, Inc.

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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6007–6017.

Seon Gyeom Kim, Jae Young Choi, Ryan Rossi, Eunyee Koh, and Tak Yeon Lee. 2025. Chart-to-experience: Benchmarking multimodal llms for predicting experiential impact of charts. In *2025 IEEE 18th Pacific Visualization Conference (PacificVis)*, pages 340–345. IEEE.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353.

Reuben A Luera, Ryan Rossi, Franck Dernoncourt, Samyadeep Basu, Sungchul Kim, Subhojyoti Mukherjee, Puneet Mathur, Ruiyi Zhang, Jihyung Kil, Nedim Lipka, et al. 2025. Mllm as a ui judge: Benchmarking multimodal llms for predicting human perception of user interfaces. *arXiv preprint arXiv:2510.08783*.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047.

Yulin Pan, Xiangteng He, Chaojie Mao, Zhen Han, Zeyinzi Jiang, Jingfeng Zhang, and Yu Liu. 2025. Icebench: A unified and comprehensive benchmark for image creating and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16586–16596.

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. 2025. Pico-banana-400k: A large-scale dataset for text-guided image editing. *arXiv preprint arXiv:2510.19808*.

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

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

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2024. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8543–8552.

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

Ron Yosef, Yonatan Bitton, Dani Lischinski, and Moran Yanuka. 2025. Editinspector: A benchmark for evaluation of text-guided image edits. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29503–29530.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2023a. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. 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.

Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. 2023b. 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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Appendix

A Problem Motivation

The purpose of this section is to further explain why traditional metrics are structurally misaligned with goals of instruction-driven image editing. Refer to table 4 for a detailed list of the traditional metrics used in our study.

A central limitation of the traditional metrics is the fundamental mismatch between what they are designed to measure and the goal of image editing. Pixel-level metrics such as PSNR and SSIM explicitly reward similarity to the original image, whereas successful editing is defined by the quality of the intentional differences. As a result, high-quality edits that introduce desirable, instruction-aligned changes are penalized as if they were distortions, causing these metrics to incorrectly classify valid edits as errors.

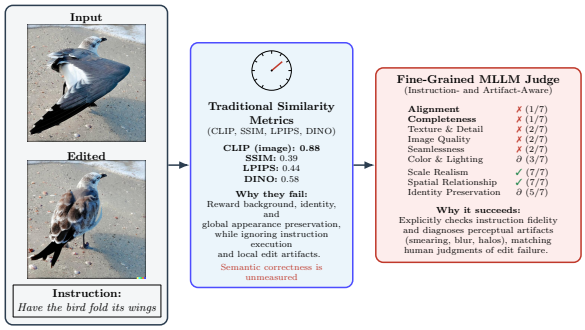

Figure 4: Similarity-based metrics reward global appearance and identity preservation, while fine-grained MLLM judges evaluate instruction fidelity and local edit quality, correctly diagnosing semantic edit failures.

Traditional metrics also fail because they conflate pixel similarity with semantic correctness. In tasks requiring substantial visual changes—such as replacing an object ("replace the dog with a cat")—metrics like PSNR and SSIM heavily penalize the large pixel differences that naturally arise from a successful transformation.

Finally, metrics like CLIPScore, while capturing coarse semantic alignment, provide no mechanism for evaluating instructional fidelity. For a prompt like "make the sky brighter," a high CLIPScore may simply reflect the co-occurrence of "sky" and "bright," regardless of whether the model brightened the correct region or overexposed unrelated areas.

B Additional Related Work

This appendix provides a more detailed discussion of prior image editing paradigms and the use of large multimodal models as evaluators.

B.1 Agentic and Tool-Based Image Editing

Agentic image editing systems decompose a complex user instruction into a sequence of specialized, interpretable operations, rather than producing the edited image in a single generative step. Recent advances in multimodal large language models (MLLMs) have enabled this paradigm by allowing models to reason over an instruction and orchestrate external tools such as segmentation, inpainting, object detection, color adjustment, and style transfer. Representative works such as InstructPix2Pix (Brooks et al., 2023) and MagicBrush (Zhang et al., 2023a) demonstrate that stepwise editing pipelines can yield localized, controllable edits and improve interpretability by making intermediate decisions explicit.

This agentic paradigm is particularly well-suited for precision-critical scenarios, where edits must be tightly constrained to specific regions of the image. By explicitly separating localization, modification, and blending, agentic systems can reduce unintended global changes and better preserve non-target regions. However, these benefits come at the cost of increased computational overhead, longer inference times, and added system complexity due to tool orchestration and error propagation across steps.

B.2 Direct Generative Image Editing

In contrast, direct image editing methods leverage a single forward pass of a generative model conditioned on the input image and a textual instruction to produce the edited output. Approaches such as Stable Diffusion `img2img` (Rombach et al., 2022), DALL·E 2 inpainting (Ramesh et al., 2022), and Imagen Editor (Saharia et al., 2022) have demonstrated strong performance for stylistic and scene-wide transformations. These models benefit from efficiency, visual coherence, and ease of deployment, making them attractive for creative and consumer-facing applications.

However, direct methods often struggle with controllability. Localized edits can inadvertently trigger global re-rendering, altering background appearance, identity features, or scene composition beyond the user's intent. For example, modifying

the color of a single object may unintentionally affect lighting, texture, or nearby objects. These failure modes directly motivate evaluation criteria related to image preservation and edit localization, which are difficult to capture using traditional similarity-based metrics.

B.3 MLLMs as Judges for Image Editing

Human evaluation remains the gold standard for assessing image editing quality, but it is costly, slow, and difficult to scale. This has motivated a growing body of work on *LLM-as-a-Judge* paradigms, where large language or multimodal models are used to approximate human preferences. Prior studies have applied LLMs and MLLMs as judges across a range of subjective evaluation tasks, including interface analysis, chart understanding, and text generation (Luera et al., 2025; Kim et al., 2025, 2024; Fu et al., 2024; Zheng et al., 2023).

More recent efforts have explored MLLM-based judges specifically for image editing (Fu et al., 2024; Hsu et al., 2023; Kim et al., 2024), demonstrating that MLLMs can reason about high-level semantics and user intent beyond pixel-level similarity. However, these approaches are not designed around the specific requirements of instruction-guided image editing as formalized in this work. In particular, they do not provide factorized evaluation aligned with the competing objectives of preserving non-edited content, producing technically realistic edits, and faithfully executing the instruction.

C Traditional Metrics

Refer to Tables 16 through for detailed correlation results between our Judge and the traditional metrics.

C.1 Pixel-Level Traditional Metrics Analysis Overview

To contextualize the performance of our MLLM-based judge, we evaluate a comprehensive set of traditional automated metrics, including pixel-level, perceptual, mask-based, and semantic similarity measures, as summarized in Table 4. Pixel-level and perceptual metrics are computed under both offline and online evaluation settings. To better isolate localized edits, we further consider mask-based variants of these metrics. In addition, we assess semantic similarity using CLIP text embeddings, CLIP image embeddings, and DINO image representations. Detailed comparison results for all metrics are provided in Table 9.

Traditional pixel-level and perceptual metrics primarily measure visual similarity and fail to capture higher-level properties such as semantic correctness, instruction fidelity, and perceptual coherence. Consistent with this limitation, we observe extremely weak correlations between all traditional metrics and our fine-grained judge factors, indicating that these metrics do not reflect the nuanced aspects of editing quality that are most relevant to human judgment. Notably, correlations in the online setting are generally stronger (i.e., more negative), suggesting that the online judge applies stricter criteria when evaluating edited images.

C.1.1 L1: Absolute Pixel Difference

L1 measures the pixel-wise absolute difference between two images, where lower values indicate greater similarity. L1 scores are typically lower in the offline setting because the edited image is compared to a known ground-truth target. Among all pixel-level metrics, L1 exhibits the strongest correlation with the *Unchanged Regions* factor in both settings, as it is sensitive to unintended pixel changes in areas that should remain untouched. L1 also correlates most strongly with *Scale Realism* in the online setting. Unrealistic scaling of the edited object alters surrounding pixels, resulting in larger absolute differences that L1 detects. Despite these being the strongest relationships among pixel-level metrics, all correlations remain very weak.

C.1.2 L2: Squared Pixel Difference

L2 measures the squared pixel-wise difference between images. It is most correlated with the *Spatial Relationship* factor in both online and offline settings. Spatial relationship errors—such as misalignment, incorrect placement, or overlapping elements—produce large pixel discrepancies that L2 captures. However, the magnitude of these correlations is still small, showing that L2 does not effectively capture higher-level semantic correctness in image editing.

C.1.3 PSNR: Peak Signal-to-Noise Ratio

PSNR compares the image signal to the noise, with higher scores indicating closer similarity. PSNR shows almost no correlation with any judge factor. Its highest correlation, still close to zero, is with *Identity Preservation* in the offline setting. Because PSNR treats all deviations as noise and lacks any

Table 4: We propose a comprehensive taxonomy that summarizes the traditional image editing metrics into pixel-level fidelity, content preservation, and semantic alignment. Each metric is categorized by purpose, assumed input(s), and reference dependence. Note that while we list assumed inputs, many of these also make sense for others, including those that the online setting assumes, where ground-truth is unknown.

	Metric	Assumed Input(s)	Reference Type
PIXEL-LEVEL FIDELITY	L1 Distance (↓)	Edited image, Ground-truth image	Reference-based
	L2 Distance / MSE (↓)	Edited image, Ground-truth image	Reference-based
	PSNR (↑)	Edited image, Ground-truth image	Reference-based
	SSIM (↑)	Edited image, Ground-truth image	Reference-based
	LPIPS (↓)	Edited image, Ground-truth image	Reference-based
CONTENT PRESERVATION	Mask-SSIM (↑)	Edited image, Ground-truth image, Edit mask	Reference-based
	Mask-LPIPS (↓)	Edited image, Ground-truth image, Edit mask	Reference-based
	Background Consistency (↑)	Edited image, Original image	Reference-based
SEMANTIC ALIGNMENT	CLIPScore (↑)	Edited image, Instruction	Reference-free
	DINO Similarity (↑)	Edited image, Ground-truth image	Reference-based

semantic understanding, it fails to provide meaningful insight into editing quality.

C.1.4 SSIM: Structural Similarity Index

SSIM measures how similar two images are based on local structural information, evaluating luminance, contrast, and structural coherence using patches rather than raw pixels. SSIM demonstrates somewhat stronger correlations with judge factors tied to overall coherence, including *Global Consistency*, *Unchanged Regions*, and *Color and Lighting*. Although these correlations remain weak, SSIM outperforms basic pixel-level metrics because its patch-based structure better captures global image consistency.

C.1.5 LPIPS: Learned Perceptual Image Patch Similarity

LPIPS compares deep visual features rather than individual pixels, making it more aligned with human perceptual judgments. LPIPS shows the strongest correlations across all pixel-level metrics, particularly with the *Texture and Detail* judge factor. This occurs because LPIPS embeddings capture fine-grained visual inconsistencies, texture mismatches, and perceptual artifacts that humans naturally identify. LPIPS also responds strongly to semantic or textual editing errors that produce noticeable perceptual differences.

C.1.6 Summary

Across all pixel-level metrics, we find that none correlate strongly with judge factors, highlighting the mismatch between pixel similarity and human evaluations of editing correctness. The online judge generally exhibits stronger correlations, suggesting higher sensitivity to editing flaws. Among all metrics, LPIPS aligns most closely with human-relevant perceptual qualities, whereas simple pixel-based metrics fail to reflect the semantic, structural, and perceptual dimensions central to instruction-based image editing.

C.2 Metrics

C.2.1 Ranked-based Correlation Metrics

Rank-based and correlation metrics are used to evaluate the consistency between MLLM-generated scores and human judgments in terms of relative ordering and linear association. Specifically, Pearson's r measures linear correlation between predicted and human scores, capturing global alignment under a parametric assumption. In contrast, Spearman's ρ and Kendall's τ focus on rank consistency, making them more robust to non-linear relationships and scale differences by evaluating whether samples are ordered similarly. In addition, we include classification-oriented measures such as Precision, Recall, and F1 score to assess the model's ability to correctly identify high- and low-quality edits under thresholded decision settings, providing a complementary perspective on discrete judgment agreement.

C.2.2 Error-based Metrics

Error-based metrics quantify the absolute discrepancy between MLLM predictions and human annotations at the score level. Mean Squared Error (MSE) penalizes larger deviations more heavily, making it sensitive to outliers and emphasizing substantial prediction errors, while Mean Absolute Error (MAE) provides a more robust measure of average deviation by treating all errors uniformly. We further report Accuracy under a tolerance-based criterion, measuring the proportion of cases where predicted scores fall within an acceptable range of human ratings. Together, these metrics capture both the magnitude and acceptability of prediction errors, offering a direct assessment of numerical

alignment between model outputs and human judgments.

C.2.3 Pairwise Accuracy

Let there be a set of n edited images (for specific edit type), and two scoring functions f_A and f_B (e.g., MLLM-as-a-Judge and a traditional metric like LPIPS). For every unordered pair of distinct images (i, j), we determine whether both scoring functions agree on their relative ordering. The *pairwise accuracy* between A and B is defined as the proportion of all item pairs for which the two systems produce consistent pairwise rankings. Formally,

$$\frac{1}{N} \sum_{i<j} \mathbf{1}\big[\mathrm{sign}\big(f_A(i) - f_A(j)\big) = \mathrm{sign}\big(f_B(i) - f_B(j)\big)\big] \tag{1}$$

where $N = \frac{n(n-1)}{2}$ is the total number of unique image pairs, and $\mathbf{1}[\cdot]$ denotes the indicator function, which equals 1 when the condition is true and 0 otherwise. Pairs where either $f(i) = f(j)$ are typically excluded to avoid ambiguity due to ties.

The resulting value lies in the range $[0, 1]$:

- 1.0 indicates perfect agreement in all pairwise comparisons,
- 0.5 indicates random agreement (no correlation),
- 0.0 represents complete disagreement (inverse ranking).

C.3 Semantic Metrics

Across the online evaluation regime, semantic similarity metrics based on CLIP and DINO exhibit only limited explanatory power for our factorized judge scores. While **DINO Image** consistently emerges as the strongest among the traditional baselines—achieving the lowest predictive error and the highest correlation values (Table 9). **CLIP Image** also offers positive alignment but with noticeably weaker associations, and **CLIP Text** has a marginal positive alignment only with task counting, underscoring the inadequacy of text-conditioned embeddings for capturing structural integrity, identity coherence, or viewpoint realism in edited imagery.

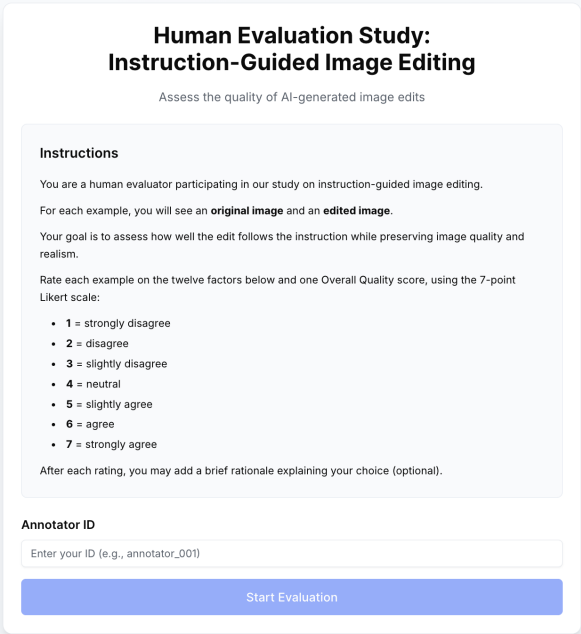

Figure 5: Human Evaluation Study UI for Instruction-Guided Image Editing

C.4 Traditional Metrics Analysis

In the online setting, correlations between our 12 judge factors and traditional evaluation metrics remain uniformly weak, underscoring the inability of similarity-based measures to capture instruction-driven editing quality. Pixel-level metrics (L1, L2, PSNR, SSIM, LPIPS) show minimal alignment overall: L1 and L2 exhibit limited sensitivity to unintended pixel changes and spatial relationship errors, PSNR provides almost no signal, SSIM modestly reflects global coherence, and LPIPS aligns most closely with human-relevant texture and detail judgments due to its feature-level perceptual modeling (Table 19). Semantic similarity metrics based on CLIP and DINO also demonstrate limited explanatory power. DINO Image performs best for content-preservation factors such as identity and global consistency, while CLIP Image is inconsistent and CLIP Text often correlates negatively with visual quality factors. These effects are amplified in the online regime, where correlations further degrade, particularly for instruction-fidelity dimensions (Table 33), highlighting a fundamental mismatch between generic similarity metrics and human-centered editing criteria.

D Human Study

D.1 User Interface

Refer to Figure 5.

Table 5: **Human and our MLLM-as-a-Judge scores for all factors and across all edit types.** We report the average score over all image edit types in the last column and over all factors in the last row. When the difference between our judge score and human evaluation is closer than 0.5, its background is dark green. When the difference is closer than 1.0, its background is light green. The gray text is the standard deviation of the scores from which the average is computed. Judge scores are generated using the Factor-level Rubrics prompt shown in Fig. 8.

	Factor		Image Edit Types						
			Add	**Remove**	**Replace**	**Action**	**Counting**	**Relation**	**All Edits**
Image Preserv.	Unchanged Regions	Human	5.172 (0.82)	5.731 (0.75)	4.972 (1.02)	4.352 (1.06)	3.393 (0.68)	4.992 (0.43)	4.769 (0.74)
		Our Judge	6.778 (0.42)	**6.000** (1.14)	6.667 (0.47)	5.826 (1.01)	**3.600** (2.06)	**5.167** (1.67)	**5.673** (1.07)
	Global Consistency	Human	5.602 (0.70)	5.982 (0.61)	5.551 (0.90)	4.769 (0.87)	5.243 (1.13)	5.444 (0.39)	5.432 (0.37)
		Our Judge	6.778 (0.42)	**6.471** (0.88)	6.667 (0.67)	6.348 (0.70)	**5.600** (1.36)	**6.167** (1.21)	**6.338** (0.39)
	Identity Preservation	Human	5.613 (0.62)	5.913 (0.82)	5.625 (0.84)	4.871 (1.12)	4.227 (1.12)	5.714 (0.20)	5.327 (0.59)
		Our Judge	7.000 (0.00)	**6.529** (0.95)	**6.556** (1.38)	6.957 (0.20)	6.400 (0.80)	**6.667** (0.75)	6.685 (0.22)
Edit Quality	Scale Realism	Human	5.276 (0.95)	6.286 (0.54)	5.865 (0.80)	5.984 (0.61)	5.510 (0.68)	6.033 (0.57)	5.826 (0.34)
		Our Judge	6.667 (0.94)	**6.853** (0.60)	6.889 (0.31)	7.000 (0.00)	7.000 (0.00)	**6.333** (1.11)	**6.790** (0.23)
	Spatial Relationship	Human	5.561 (0.79)	6.225 (0.50)	5.890 (0.63)	5.948 (0.74)	4.650 (1.19)	5.728 (0.63)	5.667 (0.50)
		Our Judge	6.667 (0.94)	**6.529** (1.06)	6.944 (0.23)	**6.609** (0.71)	**5.200** (1.17)	**6.500** (1.12)	**6.408** (0.56)
	Texture and Detail	Human	5.504 (0.56)	5.844 (0.61)	5.483 (0.80)	5.643 (0.84)	5.157 (0.90)	5.639 (0.59)	5.545 (0.21)
		Our Judge	**6.222** (1.03)	**5.794** (0.80)	**6.167** (0.37)	**6.217** (0.66)	6.200 (1.17)	**5.833** (1.07)	**6.072** (0.18)
	Image Quality	Human	5.569 (0.54)	6.048 (0.66)	5.513 (0.71)	5.947 (0.71)	5.247 (0.53)	5.683 (0.75)	5.668 (0.27)
		Our Judge	6.778 (0.42)	**6.412** (0.60)	6.889 (0.31)	**6.652** (0.56)	7.000 (0.00)	**6.500** (1.12)	6.705 (0.21)
	Color and Lighting	Human	5.515 (0.75)	5.855 (0.71)	5.442 (0.86)	5.549 (0.72)	5.403 (0.83)	5.553 (0.64)	5.553 (0.15)
		Our Judge	**6.444** (0.96)	**6.500** (0.78)	6.556 (0.50)	**6.391** (0.49)	6.600 (0.80)	**6.167** (0.90)	**6.443** (0.14)
	Seamlessness	Human	5.722 (0.64)	6.101 (0.58)	5.767 (0.74)	5.598 (0.83)	5.357 (1.00)	5.578 (0.73)	5.687 (0.23)
		Our Judge	**6.444** (1.07)	**5.941** (1.06)	**6.556** (0.50)	**6.391** (0.82)	**6.000** (1.10)	**5.667** (1.70)	**6.167** (0.32)
Instruct. Fidel.	Alignment	Human	5.556 (0.49)	5.927 (0.97)	5.681 (0.65)	5.666 (1.13)	3.437 (1.40)	5.178 (0.99)	5.241 (0.84)
		Our Judge	6.889 (0.31)	**6.471** (1.09)	**6.500** (1.01)	6.739 (0.85)	5.800 (2.40)	7.000 (0.00)	6.566 (0.39)
	Completeness	Human	5.693 (0.72)	5.966 (1.17)	5.789 (0.71)	5.719 (1.14)	3.537 (1.74)	5.556 (0.66)	5.376 (0.83)
		Our Judge	6.889 (0.31)	**6.500** (1.14)	**6.389** (1.11)	6.826 (0.64)	6.000 (2.00)	6.667 (0.75)	6.545 (0.30)
	Plausibility	Human	5.209 (1.03)	6.023 (0.78)	5.743 (0.80)	5.692 (1.17)	4.917 (1.08)	5.586 (0.89)	5.528 (0.36)
		Our Judge	6.667 (0.67)	**6.618** (0.87)	6.944 (0.23)	6.870 (0.61)	7.000 (0.00)	**6.167** (1.21)	6.711 (0.28)
	Average	Human	5.499 (0.17)	5.992 (0.15)	5.610 (0.24)	5.478 (0.50)	4.673 (0.78)	5.557 (0.26)	5.652 (0.47)
		Our Judge	6.685 (0.21)	**6.385** (0.30)	6.644 (0.23)	6.569 (0.33)	6.033 (0.92)	**6.236** (0.48)	**6.479** (0.40)

D.2 Human Judge Evaluation Metrics

Refer to Table 2, Table 5 and Table 6

D.3 Evaluation Metrics

Mean and Standard Deviation. For each factor f and edit type e, we compute the mean score across all human evaluators:

$$\mu_{f,e} = \frac{1}{N}\sum_{i=1}^{N} s_{i,f,e} \tag{2}$$

where $s_{i,f,e}$ is the score given by evaluator i for factor f on edit type e, and N is the number of evaluators. The standard deviation is:

$$\sigma_{f,e} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(s_{i,f,e} - \mu_{f,e})^2} \tag{3}$$

Intraclass Correlation Coefficient (ICC). We use ICC(2,k) (two-way random effects, average measures) to measure inter-rater reliability:

$$\mathrm{ICC}(2,k) = \frac{\mathrm{MS}_R - \mathrm{MS}_E}{\mathrm{MS}_R + (k-1)\mathrm{MS}_E + \frac{k}{n}(\mathrm{MS}_C - \mathrm{MS}_E)} \tag{4}$$

where MS_R is mean square for rows (images), MS_E is mean square error, MS_C is mean square for columns (raters), k is the number of raters per image, and n is the number of images.

D.4 Human-MLLM Alignment Metrics

Mean Squared Error (MSE).

$$\mathrm{MSE} = \frac{1}{K}\sum_{k=1}^{K}(s_k^{\text{human}} - s_k^{\text{MLLM}})^2 \tag{5}$$

Table 6: **Human and our MLLM-as-a-Judge scores for all factors and across all edit types.** We report the average score over all image edit types in the last column and over all factors in the last row. When the difference between our judge score and human evaluation is closer than 0.5, its background is dark green. When the difference is closer than 1.0, its background is light green. The gray text is the standard deviation of the scores from which the average is computed. Judge scores are generated using the Category wise, Example guided prompt shown in Fig. 9.

	Factor		Image Edit Types						
			Add	**Remove**	**Replace**	**Action**	**Counting**	**Relation**	**All Edits**
Image Preserv.	Unchanged Regions	Human	5.172 (0.82)	5.731 (0.75)	4.972 (1.02)	4.352 (1.06)	3.393 (0.68)	4.992 (0.43)	4.769 (0.74)
		Our Judge	6.333 (0.47)	**5.382** (1.19)	6.000 (0.75)	**4.739** (1.07)	**3.600** (1.62)	**4.500** (1.38)	**5.092** (0.93)
	Global Consistency	Human	5.602 (0.70)	5.982 (0.61)	5.551 (0.90)	4.769 (0.87)	5.243 (1.13)	5.444 (0.39)	5.432 (0.37)
		Our Judge	6.667 (0.47)	**6.059** (1.11)	6.389 (0.49)	5.957 (0.69)	**5.000** (1.41)	**5.500** (0.96)	**5.928** (0.55)
	Identity Preservation	Human	5.613 (0.62)	5.913 (0.82)	5.625 (0.84)	4.871 (1.12)	4.227 (1.12)	5.714 (0.20)	5.327 (0.59)
		Our Judge	7.000 (0.00)	**6.265** (1.04)	**6.278** (1.41)	6.435 (0.58)	5.200 (1.33)	**6.333** (1.11)	6.252 (0.53)
Edit Quality	Scale Realism	Human	5.276 (0.95)	6.286 (0.54)	5.865 (0.80)	5.984 (0.61)	5.510 (0.68)	6.033 (0.57)	5.826 (0.34)
		Our Judge	**6.000** (0.47)	**6.824** (0.38)	**6.389** (0.95)	**6.217** (0.51)	6.800 (0.40)	**5.833** (1.34)	**6.344** (0.37)
	Spatial Relationship	Human	5.561 (0.79)	6.225 (0.50)	5.890 (0.63)	5.948 (0.74)	4.650 (1.19)	5.728 (0.63)	5.667 (0.50)
		Our Judge	**5.778** (1.13)	**6.088** (1.20)	**6.500** (0.60)	**5.870** (0.95)	6.600 (0.49)	**5.667** (1.25)	**6.084** (0.35)
	Texture and Detail	Human	5.504 (0.56)	5.844 (0.61)	5.483 (0.80)	5.643 (0.84)	5.157 (0.90)	5.639 (0.59)	5.545 (0.21)
		Our Judge	**5.333** (1.05)	**5.529** (0.81)	**5.889** (0.66)	**5.739** (0.61)	6.000 (0.63)	**5.500** (1.26)	**5.665** (0.23)
	Image Quality	Human	5.569 (0.54)	6.048 (0.66)	5.513 (0.71)	5.947 (0.71)	5.247 (0.53)	5.683 (0.75)	5.668 (0.27)
		Our Judge	**5.778** (0.42)	**5.824** (0.57)	**6.222** (0.53)	**5.913** (0.28)	6.800 (0.40)	**5.667** (1.25)	**6.034** (0.38)
	Color and Lighting	Human	5.515 (0.75)	5.855 (0.71)	5.442 (0.86)	5.549 (0.72)	5.403 (0.83)	5.553 (0.64)	5.553 (0.15)
		Our Judge	**5.111** (0.99)	**5.735** (1.04)	**5.556** (0.83)	**5.391** (0.82)	6.800 (0.40)	**5.333** (1.25)	**5.654** (0.55)
	Seamlessness	Human	5.722 (0.64)	6.101 (0.58)	5.767 (0.74)	5.598 (0.83)	5.357 (1.00)	5.578 (0.73)	5.687 (0.23)
		Our Judge	**5.333** (1.15)	5.265 (1.27)	**5.889** (0.87)	**5.391** (0.87)	6.200 (0.40)	**4.833** (1.57)	**5.485** (0.44)
Instruct. Fidel.	Alignment	Human	5.556 (0.49)	5.927 (0.97)	5.681 (0.65)	5.666 (1.13)	3.437 (1.40)	5.178 (0.99)	5.241 (0.84)
		Our Judge	7.000 (0.00)	**6.471** (1.24)	6.556 (0.76)	6.609 (1.17)	5.800 (1.94)	6.000 (1.83)	6.406 (0.40)
	Completeness	Human	5.693 (0.72)	5.966 (1.17)	5.789 (0.71)	5.719 (1.14)	3.537 (1.74)	5.556 (0.66)	5.376 (0.83)
		Our Judge	7.000 (0.00)	**6.235** (1.54)	**6.278** (1.19)	6.609 (1.28)	6.000 (2.00)	**6.000** (1.83)	6.354 (0.35)
	Plausibility	Human	5.209 (1.03)	6.023 (0.78)	5.743 (0.80)	5.692 (1.17)	4.917 (1.08)	5.586 (0.89)	5.528 (0.36)
		Our Judge	6.889 (0.31)	**6.706** (0.75)	7.000 (0.00)	6.783 (1.02)	7.000 (0.00)	6.833 (0.37)	6.868 (0.11)
	Average	Human	5.499 (0.17)	5.992 (0.15)	5.610 (0.24)	5.478 (0.50)	4.673 (0.78)	5.557 (0.26)	5.652 (0.47)
		Our Judge	**6.185** (0.69)	**6.032** (0.48)	**6.245** (0.36)	**5.971** (0.58)	5.983 (0.95)	**5.667** (0.60)	**6.046** (0.57)

where K is the number of evaluations, s_k^{human} is the average human score for evaluation k, and s_k^{MLLM} is the MLLM judge score.

Mean Absolute Error (MAE).

$$\text{MAE} = \frac{1}{K}\sum_{k=1}^{K}|s_k^{\text{human}} - s_k^{\text{MLLM}}| \tag{6}$$

Pearson Correlation Coefficient.

$$r = \frac{\sum_{k=1}^{K}(h_k - \bar{h})(m_k - \bar{m})}{\sqrt{\sum_{k=1}^{K}(h_k - \bar{h})^2}\sqrt{\sum_{k=1}^{K}(m_k - \bar{m})^2}} \tag{7}$$

where h_k and m_k denote human and MLLM scores, respectively.

Spearman's Rank Correlation (ρ). Spearman's ρ is computed as the Pearson correlation coefficient applied to the rank-transformed scores.

Kendall's Tau (τ).

$$\tau = \frac{n_c - n_d}{\frac{1}{2}K(K-1)} \tag{8}$$

where n_c is the number of concordant pairs and n_d is the number of discordant pairs among all $\frac{1}{2}K(K-1)$ possible pairs of evaluations.

D.5 Data Storage Format

For reproducibility, we provide the exact data structure of our released benchmark dataset. Each evaluation record contains:

- `participant_id`: Anonymized evaluator identifier (string)
- `image_id`: Unique image identifier (string)

- `edit_type`: One of {Add, Remove, Replace, Action, Counting, Relation}
- `factor_scores`: Dictionary mapping each of the 12 factors to scores (1-7)
- `overall_score`: Overall quality rating (1-7)
- `timestamp_start`: Evaluation start time (ISO 8601 format)
- `timestamp_end`: Evaluation end time (ISO 8601 format)
- `annotator_id`: Participant-provided identifier (string)

Data are provided in both CSV (flat format with one row per factor per evaluation) and JSONL (nested format with one JSON object per evaluation) for convenience.

E Human vs Judge vs Traditional Agreement Metrics

E.1 Pairwise and Pointwise results

Table 7 provides detailed pointwise and pairwise agreement metrics between human evaluators and the MLLM judge across image edit types. We report error-based metrics (MSE, MAE), ACC, ACC$\pm$1, and rank correlation measures (Pearson, Spearman, Kendall's τ), and derived pairwise preference agreement statistics. For ACC, it is defined that the score for our judge and human evaluation must match each other. For ACC$\pm$1 it means the absolute difference between human evaluation and our judge is within 1.

Table 8 derived pairwise preference agreement. For each factor. Pairwise accuracy measures how often the judge recovers human preferences between image pairs.

E.2 Key Findings

For this analysis, we refer to Table 9.

MLLM judge exhibits strongest alignment with human evaluations. Across all edit types, the MLLM judge achieves an average score of 0.785 ± 0.08, closely matching and slightly exceeding the human average (0.781 ± 0.08). This level of agreement is consistently higher than that of any traditional evaluation metric, indicating that the judge captures human-aligned notions of instruction fidelity and semantic correctness that are not accessible to low-level similarity measures.

Pixel-level similarity metrics fail to reflect human judgment. Error-based and structural metrics such as L1 (0.402 ± 0.22), L2 (0.296 ± 0.22), SSIM (0.499 ± 0.24), Mask SSIM (0.499 ± 0.24), and PSNR (0.429 ± 0.23) exhibit weak alignment with human scores across all edit types. These metrics primarily assess reconstruction fidelity and are largely insensitive to whether the semantic intent of the instruction has been satisfied, leading to poor correspondence with human evaluation.

Semantic similarity metrics capture partial alignment. Feature-based metrics show improved performance relative to pixel-level measures. CLIP Image Norm (0.912 ± 0.04), CLIP Text Norm (0.593 ± 0.03), and DINO Image Norm (0.778 ± 0.19) demonstrate moderate alignment with human judgments, suggesting that semantic representations encode some aspects of instruction compliance. However, these metrics remain insufficient for reliably assessing complex edits, as they do not explicitly model edit intent, instruction grounding, or fine-grained relational correctness.

Consistency across edit types. Traditional metrics exhibit substantial variability across edit categories, particularly degrading on Action, Counting, and Relation edits. In contrast, the MLLM judge maintains comparatively stable performance across edit types, with performance drops primarily observed for inherently ambiguous cases such as Counting, where even human agreement is lower.

Implications. Overall, these results demonstrate that MLLM-based judges provide a more human-aligned evaluation signal than traditional metrics. While pixel-level and semantic similarity measures capture aspects of visual fidelity and content overlap, they fail to assess instruction satisfaction. By explicitly reasoning over intent, semantics, and contextual consistency, the MLLM judge consistently outperforms all traditional metrics in similarity to human evaluation.

E.3 Error Metrics for All Prompts

Table 7 reports a meta-evaluation of three MLLM-as-a-judge prompting strategies by comparing their agreement with human annotations across 12 fine-grained evaluation factors spanning Image Preservation, Edit Quality, and Instruction Fidelity. We report both error-based metrics (MSE, MAE; lower is better) and correlation-based metrics (Pearson, Spearman, Kendall's τ; higher is better), en-

Table 7: We compare three implementations of our fine-grained MLLM judges against human annotations across our 12 evaluation factors. Agreement is measured using error-based metrics (MSE, MAE; lower is better) and correlation-based metrics (Pearson, Spearman, Kendall's τ). For Pearson, Spearman and Kendall's τ, values larger or equal to 0.25 are bolded.

Factor	Evaluator	MSE ↓	MAE ↓	ACC ↑	ACC±1 ↑	Pearson ↑	Spearman ↑	Kendall ↑
Unchanged Regions	**Main** (Fig. 7)	2.787	1.216	0.189	0.695	0.238 (0.020)	0.130 (0.210)	0.103 (0.226)
	Factor-level Rubrics (Fig. 8)	3.050	1.258	0.168	0.695	**0.357** (<0.001)	0.214 (0.037)	0.178 (0.036)
	Category wise, Example guided (Fig. 9)	2.261	1.111	0.200	0.716	**0.423** (<0.001)	**0.384** (<0.001)	**0.310** (<0.001)
Global Consistency	**Main** (Fig. 7)	1.813	0.963	0.242	0.779	0.133 (0.200)	0.021 (0.843)	0.019 (0.832)
	Factor-level Rubrics (Fig. 8)	2.171	1.100	0.168	0.737	0.188 (0.069)	0.134 (0.196)	0.118 (0.182)
	Category wise, Example guided (Fig. 9)	2.003	0.995	0.221	0.811	0.162 (0.116)	0.191 (0.064)	0.165 (0.055)
Identity Preservation	**Main** (Fig. 7)	2.987	1.153	0.326	0.674	-0.067 (0.517)	0.006 (0.956)	0.001 (0.988)
	Factor-level Rubrics (Fig. 8)	3.345	1.289	0.242	0.653	-0.012 (0.911)	-0.049 (0.639)	-0.040 (0.649)
	Category wise, Example guided (Fig. 9)	2.713	1.163	0.242	0.705	0.145 (0.160)	0.240 (0.019)	0.201 (0.022)
Scale Realism	**Main** (Fig. 7)	1.155	0.679	0.421	0.874	0.225 (0.028)	**0.264** (0.010)	0.237 (0.010)
	Factor-level Rubrics (Fig. 8)	1.334	0.679	0.495	0.811	0.047 (0.648)	0.030 (0.769)	0.028 (0.763)
	Category wise, Example guided (Fig. 9)	0.924	0.626	0.421	0.905	**0.337** (<0.001)	**0.300** (0.003)	**0.269** (0.004)
Spatial Relationship	**Main** (Fig. 7)	1.047	0.684	0.389	0.874	**0.355** (<0.001)	0.236 (0.021)	0.211 (0.019)
	Factor-level Rubrics (Fig. 8)	1.226	0.695	0.421	0.853	**0.335** (<0.001)	**0.261** (0.011)	0.235 (0.010)
	Category wise, Example guided (Fig. 9)	1.847	0.905	0.316	0.832	0.073 (0.484)	0.105 (0.312)	0.094 (0.295)
Texture and Detail	**Main** (Fig. 7)	1.292	0.816	0.274	0.832	-0.044 (0.674)	-0.036 (0.726)	-0.030 (0.732)
	Factor-level Rubrics (Fig. 8)	1.345	0.868	0.221	0.832	0.081 (0.435)	-0.012 (0.907)	-0.007 (0.935)
	Category wise, Example guided (Fig. 9)	1.450	0.953	0.200	0.747	0.094 (0.363)	0.029 (0.779)	0.026 (0.764)
Image Quality	**Main** (Fig. 7)	0.871	0.668	0.326	0.895	0.129 (0.211)	0.123 (0.235)	0.111 (0.223)
	Factor-level Rubrics (Fig. 8)	1.018	0.700	0.326	0.884	0.201 (0.051)	0.090 (0.384)	0.082 (0.371)
	Category wise, Example guided (Fig. 9)	0.934	0.732	0.284	0.863	0.164 (0.113)	0.093 (0.369)	0.082 (0.363)
Color and Lighting	**Main** (Fig. 7)	1.655	0.911	0.295	0.779	0.058 (0.577)	0.079 (0.446)	0.069 (0.430)
	Factor-level Rubrics (Fig. 8)	1.571	0.932	0.242	0.789	**0.291** (0.004)	0.224 (0.029)	0.193 (0.031)
	Category wise, Example guided (Fig. 9)	1.897	0.995	0.284	0.726	0.131 (0.205)	0.137 (0.187)	0.115 (0.174)
Seamlessness	**Main** (Fig. 7)	1.239	0.774	0.295	0.832	0.202 (0.049)	0.222 (0.030)	0.191 (0.029)
	Factor-level Rubrics (Fig. 8)	1.566	0.879	0.253	0.789	0.167 (0.106)	0.083 (0.424)	0.070 (0.420)
	Category wise, Example guided (Fig. 9)	2.218	1.100	0.200	0.695	0.177 (0.086)	0.205 (0.046)	0.168 (0.048)
Alignment	**Main** (Fig. 7)	2.237	0.979	0.379	0.737	**0.459** (<0.001)	**0.312** (0.002)	**0.272** (0.002)
	Factor-level Rubrics (Fig. 8)	2.532	1.032	0.379	0.716	**0.377** (<0.001)	**0.275** (0.007)	0.243 (0.007)
	Category wise, Example guided (Fig. 9)	1.858	0.884	0.400	0.758	**0.593** (<0.001)	**0.483** (<0.001)	**0.422** (<0.001)
Completeness	**Main** (Fig. 7)	2.147	0.947	0.389	0.758	**0.464** (<0.001)	**0.297** (0.003)	**0.258** (0.004)
	Factor-level Rubrics (Fig. 8)	2.221	0.905	0.421	0.789	**0.434** (<0.001)	**0.304** (0.003)	**0.273** (0.003)
	Category wise, Example guided (Fig. 9)	1.895	0.832	0.442	0.800	**0.574** (<0.001)	**0.392** (<0.001)	**0.348** (<0.001)
Plausibility	**Main** (Fig. 7)	1.774	0.800	0.432	0.779	**0.278** (0.006)	0.165 (0.111)	0.147 (0.107)
	Factor-level Rubrics (Fig. 8)	1.932	0.842	0.432	0.747	**0.256** (0.012)	0.168 (0.104)	0.149 (0.101)
	Category wise, Example guided (Fig. 9)	1.953	0.853	0.421	0.758	**0.288** (0.005)	0.134 (0.194)	0.120 (0.190)
All	**Main** (Fig. 7)	1.750	0.882	0.330	0.792	0.249 (<0.001)	0.206 (<0.001)	0.177 (<0.001)
	Factor-level Rubrics (Fig. 8)	1.943	0.932	0.314	0.775	**0.275** (<0.001)	0.195 (<0.001)	0.170 (<0.001)
	Category wise, Example guided (Fig. 9)	1.829	0.929	0.303	0.776	**0.315** (<0.001)	**0.267** (<0.001)	0.224 (<0.001)

abling a comprehensive assessment of how different prompting designs align with human judgment.

In contrast, Main consistently outperforms Factor-level Rubrics across most factors in Table 7, exhibiting lower error (MSE and MAE) and stronger correlations with human judgments for a broad range of evaluation dimensions. In particular, Main achieves higher Pearson, Spearman, and Kendall correlations for all Instruction Fidelity factors, including Alignment, Completeness, and Plausibility, indicating a closer match to human relative preferences. Meanwhile, although Factor-level Rubrics introduces explicit scoring rules for each factor, its more constrained evaluation scheme appears to limit alignment with human annotations, resulting in weaker correlations and higher variance in many cases.

In contrast, Main generally exhibits lower correlation and higher error, indicating that jointly evaluating all factors within a single prompt can obscure nuanced distinctions between evaluation dimensions. Factor-level Rubrics shows intermediate performance, often improving upon Main while remaining slightly less correlated with human judgments than Category wise, Example guided prompt. This trend suggests that while a unified prompt offers practical advantages in consistency and simplicity, it may trade off some fine-grained alignment relative to factor-specific prompting.

Performance also varies systematically across evaluation categories. All judges demonstrate stronger agreement with humans on semantic related factors such as Alignment, Completeness, and Plausibility, while exhibiting weaker alignment on Image Preservation and Edit Quality factors.

Overall, these results indicate that prompt de-

Table 8: Accuracy in predicting pairwise human preferences over different prompts. The MLLM-as-a-Judge Model selected here is GPT-5-mini. for all factors and prompts. The last column shows the overall accuracy, weighted average over all factors. The highest accuracy, for both individual factors and overall, is reported in bold. For the name mapping, the order of the factors are the same with the order in Table 7. For this table, we only select image edit pairs with human evaluation score difference larger than 2 to calculate pairwise acc.

Evaluator	**UR**	**GC**	**IP**	**SR**	**SP**	**TD**	**IQ**	**CL**	**SM**	**AL**	**CP**	**PL**	**All**
Main (Fig. 7)	0.45	0.47	0.26	0.21	**0.66**	0.39	**0.44**	0.45	**0.48**	0.47	0.58	**0.35**	0.44
Factor-level Rubrics (Fig. 8)	0.52	**0.55**	0.16	0.00	0.66	**0.50**	0.35	**0.56**	0.48	0.36	0.41	0.27	0.40
Category wise, Example guided (Fig. 9)	**0.64**	0.43	**0.50**	**0.70**	0.31	0.40	0.40	0.37	0.39	**0.67**	**0.67**	0.28	**0.52**

Table 9: **Traditional metrics fail to match human evaluation, while the MLLM judge aligns closely across edit types.** We report normalized mean ± standard deviation scores for human judgments, the MLLM judge, and common traditional image evaluation metrics across six instruction-guided image edit types (Add, Remove, Replace, Action, Counting, Relation) and their aggregate (All Edits). The prompt used for MLLM Judge is Main (Figure 7). Pixel-level error metrics (L1, L2) and structural similarity measures (SSIM, Mask SSIM, PSNR) show consistently weak alignment with human judgments, while feature-based semantic metrics (CLIP Image, CLIP Text, DINO) capture only partial agreement. In contrast, the MLLM judge closely tracks human scores across all edit types, highlighting its ability to assess instruction fidelity and semantic correctness beyond visual similarity. Higher values indicate stronger agreement for similarity-based metrics, while lower values indicate better performance for error-based metrics (L1, L2). Shading denotes closeness to human judgments, with darker green indicating stronger alignment.

Metric	Add	Remove	Replace	Action	Counting	Relation	All Edits
Human Avg	0.786 (0.02)	0.856 (0.02)	0.801 (0.03)	0.783 (0.07)	0.668 (0.11)	0.794 (0.04)	0.781 (0.08)
Judge Avg	0.799 (0.07)	0.798 (0.07)	0.779 (0.09)	0.801 (0.09)	0.731 (0.08)	0.804 (0.04)	0.785 (0.08)
Background Consistency	0.499 (0.27)	0.500 (0.22)	0.494 (0.22)	0.473 (0.28)	0.554 (0.28)	0.504 (0.27)	0.499 (0.24)
Clip Image Norm	0.893 (0.04)	0.915 (0.04)	0.906 (0.04)	0.914 (0.03)	0.925 (0.03)	0.922 (0.03)	0.912 (0.04)
Clip Text Norm	0.612 (0.02)	0.571 (0.03)	0.605 (0.03)	0.619 (0.02)	0.619 (0.02)	0.630 (0.02)	0.593 (0.03)
Dino Image Norm	0.671 (0.25)	0.794 (0.20)	0.755 (0.21)	0.808 (0.30)	0.809 (0.07)	0.805 (0.15)	0.778 (0.19)
L1 Error	0.390 (0.24)	0.410 (0.20)	0.345 (0.18)	0.545 (0.28)	0.382 (0.27)	0.380 (0.21)	0.402 (0.22)
L2 Error	0.323 (0.24)	0.294 (0.22)	0.234 (0.18)	0.412 (0.23)	0.294 (0.26)	0.332 (0.23)	0.296 (0.22)
Lpips	0.550 (0.26)	0.498 (0.19)	0.525 (0.20)	0.404 (0.23)	0.526 (0.28)	0.475 (0.20)	0.500 (0.21)
Mask Lpips	0.550 (0.26)	0.498 (0.19)	0.525 (0.20)	0.404 (0.23)	0.526 (0.28)	0.475 (0.20)	0.500 (0.21)
Mask Ssim	0.499 (0.27)	0.500 (0.22)	0.494 (0.22)	0.473 (0.28)	0.554 (0.28)	0.504 (0.27)	0.499 (0.24)
Psnr	0.485 (0.27)	0.378 (0.21)	0.499 (0.20)	0.411 (0.23)	0.515 (0.28)	0.432 (0.24)	0.429 (0.23)
Ssim	0.499 (0.27)	0.500 (0.22)	0.494 (0.22)	0.473 (0.28)	0.554 (0.28)	0.504 (0.27)	0.499 (0.24)

sign plays a critical role in MLLM-based evaluation. While Category wise, Example guided prompting yields the highest alignment with human judgments, the Main evaluator used in this work provides a competitive and scalable alternative, balancing evaluation quality with practical deployment considerations.

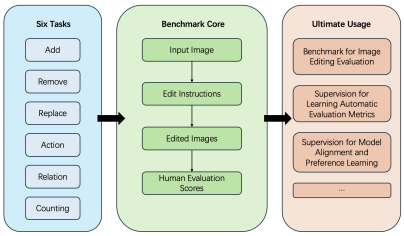

Figure 6: Structure of our benchmark and usage examples.

F Additional Benchmark Details

In this section, we provide detailed description and discussion of our proposed benchmark.

F.1 Benchmark Construction

We construct our benchmark based on the HumanEdit dataset. Specifically, for each input image and its corresponding edit instruction in HumanEdit, we use GPT-image to generate a new edited image. The original edited image provided by HumanEdit is treated as the ground-truth image. In the online setting, we evaluate the tuple (input image, edit instruction, edited image) by collecting human ratings across our designed 12 judging factors, and we additionally obtain scores from both MLLM-as-a-Judge and traditional

image quality metrics. In the offline setting, we evaluate the tuple
(edited image, edit instruction, ground truth image)
using MLLM-as-a-Judge and traditional metrics, without human involvement.

F.2 Image Preparation

All images were standardized to consistent resolution and format to ensure uniform presentation quality across evaluations. Original and edited images were presented at identical dimensions to facilitate direct visual comparison.

F.3 Benchmark Tasks

Constructed on HumanEdit, our benchmark covers six image editing tasks, including Add, Remove, Replace, Action, Relation, and Counting. The detailed definition for each edit task is listed below:

- **Add**: Add an object to the original image.
- **Remove**: Removing certain objects from an image, typically those that are more prominent or easily distinguishable.
- **Replace**: Modify the type of an object, change a part of an object, or alter its shape.
- **Counting**: Alter the number of objects in the image.
- **Action**: Alter the subject's action if the subject is a specific organism.
- **Relation**: Modifying the relationships between objects.

F.4 Benchmark Schema

For each image editing task, our benchmark contains multiple image editing instances, each consisting of an input image, an edit instruction, and the corresponding edited image generated by GPT-image. For every edited result, the benchmark provides a comprehensive set of evaluations, including traditional metric scores, five human evaluation scores, and scores produced by our MLLM-based judge, along with concise textual justifications explaining each judgment. This rich and fine-grained annotation enables systematic analysis of image editing quality across diverse tasks and evaluation perspectives.

F.5 Use Cases

The dataset serves as a benchmark for evaluating image editing models using human judgments as ground truth. It enables direct assessment of instruction following, semantic correctness, and contextual consistency, particularly in semantic editing scenarios where traditional metrics such as PSNR, SSIM, and LPIPS fail to reflect true edit quality.

The human ratings further provide supervision for learning and validating automatic evaluation metrics, including MLLM-based judges. This allows systematic analysis of metric–human alignment and supports the identification of key evaluation factors that drive accurate assessment of image editing quality.

Beyond evaluation, the dataset can be leveraged for model alignment with human preferences. Human scores or preference signals can be used to train reward models or support preference-based learning, encouraging image editing models to produce outputs that better match human expectations and editing intent.

The inclusion of complete editing context—original image, instruction, edited result, and human score—also enables fine-grained error analysis. This facilitates the study of failure modes related to instruction understanding, semantic consistency, and unintended visual artifacts.

Finally, the dataset supports systematic analysis of the gap between automatic metrics and human judgment. By comparing human ratings with metric scores, it reveals cases where existing evaluation methods overestimate or underestimate edit quality, motivating the need for more human-aligned, reasoning-based evaluation frameworks.

F.6 How to Use

Access to the full set of prompt templates, dataset loaders and evaluation code is available at our GitHub repository: https://github.com/mllmasajudge-anonymous/MLLM-as-a-Judge.

G Prompts

For all 12 factors, the prompt requires the Judge to assign a discrete score between 1 and 7 for each factor corresponding to the input image. In addition, for every judge factor, we supply textual examples illustrating scores of 1, 4, and 7 to help the Judge make more accurate quality assessments.

To examine whether different prompt formats influence the Judge's outputs, we designed the following three formats:

Table 10: 7-point Likert Scale Score used by the MLLM Judge and human evaluators.

Score	Description
1	Strongly Disagree
2	Disagree
3	Somewhat Disagree
4	Neither Agree nor Disagree
5	Somewhat Agree
6	Agree
7	Strongly Agree

- **Generate all factor results at once**: The prompt includes a single sentence definition for each factor. Each judging run produces scores for all factors at once. Refer to Figure 7 for more detail on the prompt's structure.
- **Produce results separately for each major category of factors**: For the three major categories Image Preservation, Edit Quality, and Instruction Fidelity, we generate one prompt for each category, covering all the factors within that category. Each prompt includes examples of when to score high or low for a given factor (Figure 9)
- **Produce results for all factors using a scoring rubric**: This prompt is a combination of the first two version. It generates all factors at once but also includes a scoring rubric, as illustrated in Table 11. Refer to Figure 8 for a more detailed illustration of the prompt's structure.

After having human evaluators try three prompt formats, we found that the first format did not provide enough information, as no scoring rubric was provided, and that the second format was too verbose - it required reading excessive amounts of text. Therefore, we ultimately chose to use the third prompt format, as it is a combination of the first two.

G.1 Generic Framework for all factors + offline setting

We adopt a general prompt as the input to the MLLM-as-a-Judge, which is designed to be applicable to both online and offline evaluation settings. The core structure and instructions of the prompt remain identical across settings, with only minor adjustments to the image input format to accommodate the corresponding inference pipelines. We present below the three prompt templates explored under the offline setting.

max totalsize=0.900.68,center

Main Prompt

ROLE: You are an expert image editing evaluator. Your evaluations must be objective, consistent, and grounded entirely in visual comparison and task intent.

CONTEXT: You are provided with three inputs:

1. Input Image – the unedited image.
   `[input image]`
2. Edited Image – the image produced after editing.
   `[edited image]`
3. Edit Instruction – a natural language description of the intended modification.
   `[text instruction]`

Your task is to evaluate how well the Edited Image aligns with the Input Image according to the Edit Instruction.

FACTORS:

1. Unchanged Regions: Did the parts of the image that were not supposed to be edited remain unchanged?
2. Global Consistency: Has the overall appearance (style, layout, and color) been preserved?
3. Identity Preservation: Do people, animals, or objects maintain their original identity and features after the edit?
4. Scale Realism: Is the scale of the edited object realistic compared to other objects in the image?
5. Spatial Relationship: Has the spatial relationship between objects been maintained?
6. Texture and Detail: Is the texture and detail in the edited region consistent with the surrounding areas?
7. Image Quality: Does the edited image avoid noise, blurring, or unnatural distortions?
8. Color and Lighting: Do the colors, shadows, and lighting of the edited region match the rest of the image?
9. Seamlessness: Does the transition between edited and non-edited regions look natural?
10. Alignment: Does the edited image align with the specific edits provided in the instructions?
11. Completeness: Were all aspects of the instruction carried out fully?
12. Plausibility: Does the result make sense in a real-world context?

EVALUATION STEPS:

1. Compare the Edited Image to the Ground Truth Image in the context of the Edit Instruction.
2. Assess how well the edited image satisfies each factor definition.
3. Assign a score between 1 and 7 (integers only) using the rubric above.
4. Provide a concise justification (10–25 words) describing what evidence supports your score.

SCORING (7-point Likert Scale):

[Refer to Table 10 for a detailed scoring rubric]

Decimal values are not allowed. Use the rubric to guide your scoring.

OUTPUT FORMAT (strict JSON):

[Refer to section G.2 for a detailed JSON output schema]

CONSTRAINTS:

1. Respond with only one JSON block.
2. The score must be an integer between 1 and 7.
3. The justifications must reference specific visible evidence (not speculation).
4. Do not restate the definition or include reasoning chains.
5. Keep the tone factual, concise, and visually grounded.

Figure 7: MLLM-as-a-Judge Image Editing Main prompt with no explicit rubric provided for each fine-grained factor, online setting.

Table 11: Rubric for Image Editing Factors (7-point Likert scale). This rubric is used in our fine-grained MLLM judge for image editing implementation shown in.

Factor	Score 1	Score 4	Score 7
Unchanged Regions	The model changed large areas unrelated to the instruction	Small artifacts exist but most regions are intact	No unintended change is visible
Global Consistency	The overall style, layout, or color scheme is drastically different	Minor inconsistencies in style or layout are present	The overall appearance is fully consistent
Identity Preservation	Core identifying features have been significantly altered or lost	Some features have changed but entities remain generally recognizable	All entities retain their distinguishing characteristics perfectly
Scale Realism	The edited object's scale is highly unrealistic or implausible	The scale is somewhat off but not jarringly unrealistic	The scale is completely realistic and proportionate
Spatial Relationship	Objects are misplaced or spatial relationships are severely disrupted	Minor spatial inconsistencies exist but overall relationships hold	All spatial relationships are perfectly maintained
Texture and Detail	Texture is notably different or detail is significantly degraded	Texture matches reasonably well with minor detail inconsistencies	Texture and detail are seamlessly consistent throughout
Image Quality	Severe noise, blurring, or distortions are present	Minor quality issues are noticeable but not severe	Image quality is excellent with no artifacts
Color and Lighting	Colors or lighting are severely mismatched with obvious inconsistencies	Colors and lighting mostly match with minor discrepancies	Colors, shadows, and lighting are perfectly harmonious
Seamlessness	Transitions are obvious with clear visible boundaries or seams	Transitions are mostly smooth with minor detectable edges	Transitions are completely seamless and undetectable
Alignment	The edit does not match the instruction or contradicts it	The edit partially matches but misses some key aspects	The edit perfectly matches all aspects of the instruction
Completeness	Major parts of the instruction were not executed	Most aspects were completed but some elements are missing	Every aspect of the instruction was fully executed
Plausibility	The result is highly implausible or violates real-world logic	The result is somewhat plausible but has noticeable oddities	The result is completely plausible and realistic

max totalsize=0.900.68,center

Factor-level Rubrics Prompt

ROLE: You are an expert image editing evaluator. Your evaluations must be objective, consistent, and grounded entirely in visual comparison and task intent.

CONTEXT: You are provided with three inputs:

1. Input Image – the unedited image.
   `[input image]`
2. Edited Image – the image produced after editing.
   `[edited image]`
3. Edit Instruction – a natural language description of the intended modification.
   `[text instruction]`

Your task is to evaluate how well the Edited Image aligns with the Input Image according to the Edit Instruction.

FACTORS:

[Refer to Table 11 for a detailed list of factors and a rubric for each factor]

EVALUATION STEPS:

1. Compare the Edited Image to the Ground Truth Image in the context of the Edit Instruction.
2. Assess how well the edited image satisfies each factor definition.
3. Assign a score between 1 and 7 (integers only) using the rubric above.
4. Provide a concise justification (10–25 words) describing what evidence supports your score.

SCORING (7-point Likert Scale):

[Refer to Table 10 for a detailed scoring rubric]

Decimal values are not allowed. Use the rubric to guide your scoring.

OUTPUT FORMAT (strict JSON):

[Refer to section G.2 for a detailed JSON output schema]

CONSTRAINTS:

1. Respond with only one JSON block.
2. The score must be an integer between 1 and 7.
3. The justifications must reference specific visible evidence (not speculation).
4. Do not restate the definition or include reasoning chains.
5. Keep the tone factual, concise, and visually grounded.

Figure 8: MLLM-as-a-Judge Image Editing Factor-level Rubrics prompt used for MLLM-as-a-Judge, online setting.

max totalsize=0.900.68,center

Category wise, Example guided Prompt

ROLE: You are an expert image editing evaluator specializing in instruction fidelity analysis. Your evaluations must be objective, consistent, and grounded entirely in assessing how well the edit follows the given instruction.

CONTEXT: You are provided with three inputs:

1. **Input Image** – the original image before any editing [input image]
2. **Edited Image** – the image produced after applying the edit instruction [edited image]
3. **Edit Instruction** – a natural language description of the intended modification [text instruction]

Your task is to evaluate how faithfully the Edited Image executes the Edit Instruction. You will assess three specific factors related to instruction fidelity.

FACTORS UNDER REVIEW:

=== FACTOR 1: ALIGNMENT ===

Definition: Evaluates whether the edited image aligns with the specific edits provided in the instructions—whether what was requested was actually done.

What to examine:

1. Parse the Edit Instruction carefully to identify all requested changes (e.g., "change the car to red" requests a color change to red).
2. Check whether each requested change is present in the Edited Image.
3. Verify accuracy of the changes: if the instruction asks for "red," is it red (not orange or pink)? If it asks for "a dog," is there a dog (not a cat)?
4. Assess specificity matching: if the instruction specifies "vintage wooden chair," does the result show a vintage wooden chair (not a modern plastic chair)?
5. Check for correct targets: if the instruction says "change the woman's hat," was the woman's hat changed (not someone else's or a different item)?
6. Evaluate whether the edit follows the instruction's intent and specific requirements.

Important: This factor focuses on WHETHER the requested changes match what was asked for. Do not evaluate if ALL parts were done (that's Completeness) or if the result is realistic (that's Plausibility).

Score high (6–7) when:

1. All requested changes accurately match the instruction's specifications.
2. Target objects/attributes are correctly identified and modified.
3. Specific details (colors, object types, attributes) align precisely with what was requested.
4. The edit clearly follows the instruction's intent.

Score low (1–3) when:

1. Requested changes are incorrect or inaccurate (wrong color, wrong object type, etc.).
2. Wrong elements were modified instead of the specified targets.
3. The edit contradicts or misinterprets the instruction.
4. Specific requirements are ignored or incorrectly executed.

EVALUATION STEPS:

1. Read the Edit Instruction carefully and identify all requested changes, targets, and specifications.
2. For each factor, systematically examine the Edited Image in relation to the instruction.
3. Compare the Edited Image to the Input Image to understand what changed.
4. Look for specific evidence relevant to each factor definition.
5. Assign a score between 1 and 7 (integers only) for each factor using the Likert scale below.
6. Provide a concise justification (15–30 words) for each factor, citing specific observable evidence.

SCORING (7-point Likert Scale):

[Refer to Table 10 for a detailed scoring rubric]

Decimal values are not allowed. Use the rubric to guide your scoring.

OUTPUT FORMAT (strict JSON):

[Refer to section G.2 for a detailed JSON output schema]

CONSTRAINTS:

1. Respond with only one JSON block containing all three factors.
2. Each score must be an integer between 1 and 7.
3. Each justification must reference specific, observable visual evidence.
4. Do not restate definitions or include reasoning chains in justifications.
5. Be precise: identify what aspects of the instruction were or were not followed.
6. Remain objective: evaluate only what is visible and what the instruction requested.
7. Keep tone factual, concise, and visually grounded.
8. Evaluate each factor independently—do not let one factor's assessment influence another.
9. For Alignment: focus on accuracy of what was done.
10. For Completeness: focus on thoroughness—whether everything was done.
11. For Plausibility: focus on real-world possibility of the result.

Figure 9: MLLM-as-a-Judge Image Editing Category wise, Example guided prompt for the alignment factor within the instruction fidelity category, separated by categories, online setting. The other factors follow the same detailed format.

G.2 JSON output

max totalsize=0.950.78,center

JSON scheme

```
{
  "image_id": "<image_identifier>",
  "offline_factor_results": {
    "unchanged_regions": {
      "score": <integer_1_to_7>,
      "justification": "<brief_justification>"
    },
    "global_consistency": {
      "score": <integer_1_to_7>,
      "justification": "<brief_justification>"
    },
    "identity_preservation": {
      "score": <integer_1_to_7>,
      "justification": "<brief_justification>"
    },
    "scale_realism": {
      "score": <integer_1_to_7>,
      "justification": "<brief_justification>"
    },
    "spatial_relationship": {
      "score": <integer_1_to_7>,
      "justification": "<brief_justification>"
    },
    "texture_and_detail": {
      "score": <integer_1_to_7>,
      "justification": "<brief_justification>"
    },
    "image_quality": {
      "score": <integer_1_to_7>,
      "justification": "<brief_justification>"
    },
    "color_and_lighting": {
      "score": <integer_1_to_7>,
      "justification": "<brief_justification>"
    },
    "seamlessness": {
      "score": <integer_1_to_7>,
      "justification": "<brief_justification>"
    },
    "alignment": {
      "score": <integer_1_to_7>,
      "justification": "<brief_justification>"
    },
    "completeness": {
      "score": <integer_1_to_7>,
      "justification": "<brief_justification>"
    },
    "plausibility": {
      "score": <integer_1_to_7>,
      "justification": "<brief_justification>"
    }
  }
}
```

H Ablation Study

In this section, we will provide detailed results on how different MLLM base model and different prompt formats may influence the MLLM-as-a-Judge's performance.

H.1 Human Evaluation Results Comparison

H.1.1 Varying MLLMs

Table 12 presents a comparative analysis of human evaluations alongside two MLLM judges, GPT and Gemini, across all factors and image edit types. Overall, both judges captured meaningful aspects of human judgment, but the GPT judge exhibited consistently stronger alignment with human evaluations. This was reflected by the larger number of green-highlighted cells for GPT, indicating smaller absolute differences from human scores across a wide range of factors.

In particular, GPT showed robust agreement with human ratings for image preservation and edit quality factors, including Unchanged Regions, Global Consistency, Spatial Relationship, Texture and Detail, Color and Lighting, and Seamlessness, where its scores closely tracked human averages across most edit types. In contrast, even though the Gemini judge perform better than GPT judge in several cases, like Plausibility&Add, Spatial Relationship&Replace, and Scale Realism&Action, the Gemini judge displayed larger deviations and higher variance in most factors, and got a weaker alignment than GPT judge in nearly all the factors results.

Taken together, these results suggested that the GPT judge provided a more reliable and human-aligned evaluation signal than Gemini, particularly when assessing fine-grained perceptual quality and overall edit correctness across diverse editing scenarios.

H.1.2 Varying Prompts

Refer to Table 2, Table 5 and Table 6 for a comprehensive comparison between human evaluations and our MLLM-as-a-Judge under different prompt designs. Overall, the Main evaluator(Figure 7) demonstrated the strongest general alignment with human evaluations across factors and edit types, as evidenced by the noticeably larger number of green-highlighted cells in these tables. Since green regions indicate small absolute differences between judge and human scores, this pattern suggested that the Main evaluator most consistently reproduced human rating behavior at the factor level.

As reported in Table 7, our Main evaluator (Fig 7) achieved consistently low MSE and MAE across a wide range of evaluation factors, by observing that MAE is less than 1 across nearly all factors indicating a close alignment with human absolute score annotations. However, it should be noticed that in some cases MAE is less than 1 but MSE is larger than 1 like for factor Global Consistency, Scale Realism, and Alignment. This means that although the judge performs well generally, the performance is not good as expected in some difficult cases. The Main evaluator also exhibited stable accuracy and tolerance-based accuracy (ACC and ACC$\pm$1) across most factors, with ACC close to 0.3 in most cases and ACC$\pm$1 reaching almost 0.9 in some factors, suggesting reliable pointwise agreement with human judgments.

Beyond error-based metrics, the Factor evalua-

Table 12: **Human vs. GPT vs. Gemini judge scores across factors and edit types.** Mean ± std 1–7 Likert scores for human annotations and two MLLM judges (GPT and Gemini). When the difference between our judge score and human evaluation is closer than 0.5, its background is dark green. When the difference is closer than 1.0, its background is light green. The prompt used in both models is Factor-level Rubrics (Figure 8)

			Image Edit Types						
	Factor		Add	Remove	Replace	Action	Counting	Relation	All Edits
Image Presery.	Unchanged Regions	Human	5.172 (0.82)	5.731 (0.75)	4.972 (1.02)	4.352 (1.06)	3.393 (0.68)	4.992 (0.43)	4.769 (0.74)
		GPT Judge	5.889 (1.20)	**5.235** (1.68)	5.778 (1.69)	5.522 (1.44)	**3.600** (1.96)	**5.167** (1.67)	**5.198** (0.76)
		Gemini Judge	**5.111** (2.69)	4.206 (2.67)	3.444 (2.39)	2.957 (2.69)	2.200 (2.40)	4.000 (3.00)	3.653 (0.93)
	Global Consistency	Human	5.602 (0.70)	5.982 (0.61)	5.551 (0.90)	4.769 (0.87)	5.243 (1.13)	5.444 (0.39)	5.432 (0.37)
		GPT Judge	**5.556** (1.07)	**5.618** (1.51)	**5.778** (1.62)	5.783 (1.06)	4.200 (1.94)	**5.500** (1.50)	**5.406** (0.55)
		Gemini Judge	**5.111** (2.69)	4.824 (2.53)	4.333 (2.40)	3.870 (2.52)	2.400 (2.33)	4.333 (2.75)	4.145 (0.87)
	Identity Preservation	Human	5.613 (0.62)	5.913 (0.82)	5.625 (0.84)	4.871 (1.12)	4.227 (1.12)	5.714 (0.20)	5.327 (0.59)
		GPT Judge	**5.444** (1.57)	**6.324** (1.51)	**5.889** (1.91)	6.696 (0.86)	5.400 (1.36)	5.167 (2.11)	**5.820** (0.54)
		Gemini Judge	**5.556** (2.27)	5.324 (2.37)	4.222 (2.64)	3.478 (2.72)	1.400 (0.80)	3.500 (2.57)	3.913 (1.38)
Edit Quality	Scale Realism	Human	5.276 (0.95)	6.286 (0.54)	5.865 (0.80)	5.984 (0.61)	5.510 (0.68)	6.033 (0.57)	5.826 (0.34)
		GPT Judge	**5.000** (1.94)	**6.471** (1.40)	6.667 (1.00)	7.000 (0.00)	**5.800** (1.47)	5.333 (1.37)	**6.045** (0.72)
		Gemini Judge	5.889 (1.97)	6.824 (1.01)	6.667 (1.37)	**6.130** (1.96)	6.800 (0.40)	**6.500** (0.76)	6.468 (0.35)
	Spatial Relationship	Human	5.561 (0.79)	6.225 (0.50)	5.890 (0.63)	5.948 (0.74)	4.650 (1.19)	5.728 (0.63)	5.667 (0.50)
		GPT Judge	4.778 (1.55)	**6.471** (1.33)	6.833 (0.50)	**6.217** (1.02)	**4.200** (1.60)	4.000 (1.91)	**5.417** (1.13)
		Gemini Judge	4.222 (2.90)	**6.235** (1.90)	**5.889** (2.16)	4.696 (2.56)	3.600 (2.33)	4.333 (2.75)	4.829 (0.93)
	Texture and Detail	Human	5.504 (0.56)	5.844 (0.61)	5.483 (0.80)	5.643 (0.84)	5.157 (0.90)	5.639 (0.59)	5.545 (0.21)
		GPT Judge	**5.222** (0.79)	5.176 (1.48)	4.611 (1.83)	**5.826** (0.87)	**5.000** (1.79)	**5.167** (1.21)	**5.167** (0.36)
		Gemini Judge	4.000 (2.40)	3.735 (2.59)	4.389 (2.38)	3.609 (2.62)	4.400 (2.33)	3.667 (2.56)	3.967 (0.33)
	Image Quality	Human	5.569 (0.54)	6.048 (0.66)	5.513 (0.71)	5.947 (0.71)	5.247 (0.53)	5.683 (0.75)	5.668 (0.27)
		GPT Judge	6.667 (0.47)	**6.235** (0.94)	6.667 (0.75)	**6.348** (0.91)	5.800 (1.60)	**6.000** (1.00)	6.286 (0.32)
		Gemini Judge	**5.556** (2.11)	5.000 (2.35)	**5.611** (2.16)	4.652 (2.46)	7.000 (0.00)	**5.500** (2.29)	**5.553** (0.73)
	Color and Lighting	Human	5.515 (0.75)	5.855 (0.71)	5.442 (0.86)	5.549 (0.72)	5.403 (0.83)	5.553 (0.64)	5.553 (0.15)
		GPT Judge	**6.000** (0.94)	**5.794** (1.21)	**5.611** (1.57)	**5.913** (1.06)	**5.200** (2.23)	**5.500** (1.12)	**5.670** (0.27)
		Gemini Judge	4.222 (2.39)	3.441 (2.52)	4.611 (2.52)	3.826 (2.51)	3.600 (2.80)	4.667 (2.21)	4.061 (0.47)
	Seamlessness	Human	5.722 (0.64)	6.101 (0.58)	5.767 (0.74)	5.598 (0.83)	5.357 (1.00)	5.578 (0.73)	5.687 (0.23)
		GPT Judge	**5.667** (0.82)	5.118 (1.79)	**5.778** (1.44)	**5.913** (1.06)	4.200 (2.04)	4.667 (1.37)	**5.224** (0.63)
		Gemini Judge	**5.333** (2.36)	4.059 (2.63)	**6.000** (2.00)	4.348 (2.53)	**5.200** (2.40)	4.167 (2.85)	4.851 (0.71)
Instruct. Fidel.	Alignment	Human	5.556 (0.49)	5.927 (0.97)	5.681 (0.65)	5.666 (1.13)	3.437 (1.40)	5.178 (0.99)	5.241 (0.84)
		GPT Judge	**5.444** (1.71)	4.706 (2.52)	3.833 (2.63)	**5.522** (2.38)	2.800 (1.60)	3.667 (2.56)	4.329 (0.99)
		Gemini Judge	3.111 (1.29)	3.559 (2.80)	2.278 (1.69)	3.217 (2.48)	4.000 (2.68)	2.667 (2.05)	3.139 (0.56)
	Completeness	Human	5.693 (0.72)	5.966 (1.17)	5.789 (0.71)	5.719 (1.14)	3.537 (1.74)	5.556 (0.66)	5.376 (0.83)
		GPT Judge	**6.111** (1.66)	4.618 (2.49)	3.833 (2.61)	**5.435** (2.37)	2.600 (1.50)	3.500 (2.36)	4.349 (1.18)
		Gemini Judge	5.000 (2.83)	4.294 (2.92)	2.778 (2.46)	3.957 (2.79)	4.600 (2.94)	3.167 (2.73)	3.966 (0.78)
	Plausibility	Human	5.209 (1.03)	6.023 (0.78)	5.743 (0.80)	5.692 (1.17)	4.917 (1.08)	5.586 (0.89)	5.528 (0.36)
		GPT Judge	6.000 (1.15)	**6.353** (1.26)	6.444 (1.01)	6.826 (0.64)	**5.200** (1.17)	**5.500** (0.76)	6.054 (0.56)
		Gemini Judge	**5.556** (2.17)	5.353 (2.58)	**6.056** (2.12)	4.391 (2.81)	7.000 (0.00)	4.000 (3.00)	**5.393** (1.00)

tor showed meaningful positive correlations with human annotations for several perceptual and structural factors (Table 2), notably Spatial Relationship, Image Quality, Color and Lighting, and Seamlessness, reflecting its ability to preserve relative ranking and preference structure in human evaluations. This behavior was further supported by the pairwise preference prediction results in Table 8, where the Factor evaluator attained strong accuracy across multiple factors and a competitive overall weighted accuracy.

A more fine-grained analysis of the pointwise and pairwise evaluation results in Table 7 and Table 8 shows that in terms of pointwise error metrics (MSE and MAE) and pairwise preference prediction accuracy, the Category wise, Example guided prompt (Figure 9) often achieved the strongest numerical performance across multiple factors, particularly for instruction-fidelity dimensions such as Alignment and Completeness. However, despite its favorable quantitative results, the Category wise, Example guided introduced a substantially longer and more complex instruction format that was impractical for human evaluators, as it imposed a higher cognitive load and increased the likelihood of annotation fatigue. In contrast, the Main evaluator mirrored the human evaluation setup in both structure and level of detail, enabling a more direct and fair comparison between human and judge assessments. Consequently, when the objective was to evaluate alignment with realistic human evaluation behavior rather than optimizing raw agreement metrics in isolation, the Main evaluator emerged as the most suitable and representative configuration.

Finally, while the Factor-level Rubrics prompt (Figure 8) underperformed relative to the other two variants in many cases while adding a detailed rubric on the base of Main evaluator, the underlying

causes of this degradation presented an interesting direction for future investigation and were therefore deferred to the ablation study. We think this phenomena is interesting and worse to dug deeper in, but we won't discuss about this much in this paper.

H.2 Traditional Metrics Results Comparison

H.2.1 Varying MLLMs

For this section, we will compare the performance of two different MLLMs, GPT-5-mini and Gemini-2.5-pro. The prompts we used is the Factor-level Rubrics prompt as can be seen in Appendix G Figure 8. The results for offline setting can be seen in Table 16, Table 17, Table 18,Table 22, Table 23, Table 24, Table 28, Table 29, and Table 30. Alternatively, the results for online setting can be seen in Table 19, Table 20, Table 21, Table 25, Table 26, Table 27, Table 31, Table 32, and Table 33

I Inter-Factor Analysis

Our framework does not assume statistical independence among factors, interpretability and diagnostic value were the primary design goals. We include a lightweight empirical analysis of inter-factor correlations. As shown in Table 13, certain factors do exhibit moderate correlation, which aligns with intuitive semantic coupling (e.g., alignment - completeness, texture detail - image quality, and unchanged region - identity preservation). However, other factors remain relatively orthogonal, and critically, models with similar overall edit quality can differ substantially on individual factors, underscoring the utility of the full taxonomy. This analysis supports the claim that the proposed factorization captures meaningful and complementary aspects of human edit quality judgments.

J Alignment between Human Annotators

Our validation does not rely on a single annotator group or unexamined subjective agreement. For each image, we collected ratings from multiple annotators and explicitly analyzed inter-annotator variability across all 12 evaluation factors. In Table 14, we computed the Mean Std, Median Std, Mean Absolute Difference (MAD) and Median Absolute Difference between annotators' scores for each factor. The results show that all the mean std and median std are below 1.5, the mean absolute differences are consistently below 2 across all categories, while the median absolute difference is 1 for most factors. This indicates that annotators generally differed by at most one point on the rating scale, demonstrating strong consistency rather than large subjective divergence.

More importantly, our annotator pool was intentionally diverse. It included participants from both academia and industry, spanning professors, researchers, and students, as well as individuals across different age groups and professional backgrounds. This diversity reduces the likelihood that the benchmark reflects the preference of a narrow or homogeneous group. Taken together, the low inter-annotator variability and heterogeneous annotator composition suggest that the collected ratings reflect stable and broadly shared evaluation standards rather than idiosyncratic bias from a specific subgroup.

K Editing Quality Check

The edited images used for comparison with ground truth were generated by GPT-Image, which is a relatively strong model and therefore produces many reasonably high-quality edits. However, we ensured that the dataset is not composed solely of high-scoring samples. As shown in Table 15, across all major edit categories and the 12 evaluation factors, at least 15% of samples in each category received scores of 4 or below from human evaluators, with some categories exceeding 30%. This confirms that the benchmark includes a substantial proportion of lower-quality edits, enabling the judge to learn both positive and negative evaluation signals.

L AI Usage

For this work, we used AI for paraphrasing and polishing our original essay. The use of tools only assists with proofreading.

M Benchmark License

Our dataset is released under a non-commercial, research-only license that permits use, redistribution, and modification for academic and internal research purposes, while prohibiting commercial use. The dataset inherits and complies with the licensing terms of HumanEdit (CC-BY 4.0), and all annotations we contribute are released under the same or a compatible license to avoid downstream ambiguity.

Table 13: **Inter-factor correlation matrix among the 12 evaluation factors.**

	Unchanged Regions	Global Consistency	Identity Preservation	Scale Realism	Spatial Relationship	Texture Detail	Image Quality	Color Lighting	Seamlessness	Alignment	Completeness	Plausibility
Unchanged Regions	1.000	0.691	0.742	0.427	0.471	0.568	0.572	0.516	0.555	0.486	0.389	0.476
Global Consistency	0.691	1.000	0.700	0.533	0.501	0.620	0.585	0.598	0.672	0.393	0.291	0.501
Identity Preservation	0.742	0.700	1.000	0.531	0.541	0.585	0.583	0.541	0.604	0.462	0.378	0.542
Scale Realism	0.427	0.533	0.531	1.000	0.762	0.670	0.655	0.645	0.692	0.464	0.395	0.631
Spatial Relationship	0.471	0.501	0.541	0.762	1.000	0.664	0.627	0.556	0.656	0.519	0.447	0.636
Texture Detail	0.568	0.620	0.585	0.670	0.664	1.000	0.778	0.713	0.733	0.425	0.353	0.589
Image Quality	0.572	0.585	0.583	0.655	0.627	0.778	1.000	0.670	0.721	0.494	0.404	0.639
Color Lighting	0.516	0.598	0.541	0.645	0.556	0.713	0.670	1.000	0.650	0.414	0.341	0.532
Seamlessness	0.555	0.672	0.604	0.692	0.656	0.733	0.721	0.650	1.000	0.461	0.385	0.719
Alignment	0.486	0.393	0.462	0.464	0.519	0.425	0.494	0.414	0.461	1.000	0.855	0.534
Completeness	0.389	0.291	0.378	0.395	0.447	0.353	0.404	0.341	0.385	0.855	1.000	0.461
Plausibility	0.476	0.501	0.542	0.631	0.636	0.589	0.639	0.532	0.719	0.534	0.461	1.000

Table 14: **Statistics of score differences between human evaluation across the 12 evaluation factors.** The table reports both dispersion statistics (standard deviation) and absolute score differences.

Factor	Mean Std	Median Std	Mean Abs Diff (MAE)	Median Abs Diff
Unchanged Regions	1.483	1.497	2.029	2.000
Global Consistency	1.293	1.356	1.744	1.000
Identity Preservation	1.347	1.344	1.836	1.000
Scale Realism	1.124	1.095	1.459	1.000
Spatial Relationship	1.121	1.020	1.455	1.000
Texture Detail	1.264	1.265	1.695	1.000
Image Quality	1.181	1.095	1.621	1.000
Color Lighting	1.245	1.299	1.710	1.000
Seamlessness	1.123	1.020	1.530	1.000
Alignment	1.170	1.000	1.575	1.000
Completeness	1.047	0.866	1.407	1.000
Plausibility	1.195	1.166	1.563	1.000

Table 15: **Proportion of scores smaller or equal to 4 from human evaluators by category**

Category	Unchanged Regions	Global Consistency	Identity Preservation	Scale Realism	Spatial Relationship	Texture and Detail	Image Quality	Color and Lighting	Seamlessness	Alignment	Completeness	Plausibility	**Overall**
Action	48.9%	39.0%	34.8%	12.8%	12.8%	19.9%	13.5%	20.6%	19.1%	23.4%	19.9%	20.6%	45.4%
Add	27.3%	18.2%	22.7%	29.5%	20.5%	22.7%	22.7%	22.7%	15.9%	18.2%	18.2%	36.4%	40.9%
Counting	63.0%	33.3%	51.9%	22.2%	37.0%	33.3%	33.3%	29.6%	25.9%	63.0%	63.0%	40.7%	59.3%
Relation	32.1%	28.6%	21.4%	10.7%	21.4%	25.0%	25.0%	28.6%	28.6%	28.6%	21.4%	21.4%	39.3%
Remove	19.2%	16.2%	18.6%	9.6%	9.6%	18.0%	13.8%	18.0%	13.8%	16.8%	15.6%	18.6%	23.4%
Replace	33.3%	21.5%	23.7%	20.4%	16.1%	24.7%	25.8%	25.8%	18.3%	20.4%	18.3%	17.2%	28.0%

Table 16: Results of Image Preservation showing five evaluation metrics for all factors using traditional metrics (L1, L2, PSNR, SSIM, LPIPS) (OFFLINE setting, refer to Figure 8 for prompt). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

	Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
IMAGE PRESERVATION	Unchanged Regions	L1 (↓)	GPT	0.320	0.504	-0.252	**-0.212**	**-0.176**
			Gemini	0.366	0.547	-0.197	-0.225	-0.180
		L2 (↓)	GPT	0.365	0.535	-0.211	-0.179	-0.138
			Gemini	0.406	0.568	-0.190	-0.158	-0.127
		PSNR (↑)	GPT	0.183	**0.333**	0.234	0.179	0.138
			Gemini	0.261	0.434	0.081	0.158	0.127
		SSIM (↑)	GPT	**0.175**	0.338	**0.264**	0.180	0.138
			Gemini	0.276	0.442	-0.004	-0.022	-0.009
		LPIPS (↓)	GPT	0.199	0.385	0.036	0.104	0.073
			Gemini	0.253	0.454	0.042	-0.044	-0.027
		Avg	GPT	0.248	0.419	0.014	0.014	0.007
			Gemini	0.312	0.489	-0.054	-0.058	-0.043
	Global Consistency	L1 (↓)	GPT	0.244	0.416	-0.026	-0.079	**-0.076**
			Gemini	0.370	0.554	-0.168	-0.251	-0.186
		L2 (↓)	GPT	0.280	0.455	0.007	-0.002	0.008
			Gemini	0.418	0.586	-0.193	-0.236	-0.162
		PSNR (↑)	GPT	0.205	0.360	-0.046	0.002	-0.008
			Gemini	0.248	0.412	0.168	0.236	0.162
		SSIM (↑)	GPT	**0.178**	**0.340**	**0.091**	**0.084**	0.060
			Gemini	0.298	0.484	-0.062	-0.041	-0.032
		LPIPS (↓)	GPT	0.182	0.359	-0.020	-0.044	-0.055
			Gemini	0.272	0.473	-0.011	-0.015	-0.009
		Avg	GPT	0.218	0.386	0.001	-0.008	-0.014
			Gemini	0.321	0.502	-0.053	-0.061	-0.045
	Identity Preservation	L1 (↓)	GPT	0.317	0.505	-0.287	-0.241	-0.181
			Gemini	0.393	0.564	-0.208	-0.188	-0.150
		L2 (↓)	GPT	0.354	0.520	-0.257	-0.226	-0.165
			Gemini	0.450	0.605	-0.244	-0.209	-0.176
		PSNR (↑)	GPT	0.176	0.340	0.246	0.226	0.165
			Gemini	0.258	0.430	0.140	0.209	0.176
		SSIM (↑)	GPT	**0.122**	**0.283**	**0.529**	**0.533**	**0.423**
			Gemini	0.292	0.467	-0.034	0.017	0.013
		LPIPS (↓)	GPT	0.248	0.428	-0.266	-0.248	-0.165
			Gemini	0.264	0.452	0.041	-0.002	0.019
		Avg	GPT	0.243	0.415	-0.007	0.009	0.015
			Gemini	0.331	0.504	-0.061	-0.035	-0.024

Table 17: Results of Edit Quality showing five evaluation metrics for all factors using traditional metrics (L1, L2, PSNR, SSIM, LPIPS) (OFFLINE setting, refer to Figure 8 for prompt). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

	Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
EDIT QUALITY	Scale Realism	L1 (↓)	GPT	0.212	0.394	0.077	0.007	0.015
			Gemini	0.440	0.616	-0.463	-0.461	-0.368
		L2 (↓)	GPT	0.240	0.413	0.098	0.049	0.036
			Gemini	0.514	0.675	-0.555	-0.558	-0.429
		PSNR (↑)	GPT	0.226	0.407	**-0.123**	-0.049	-0.036
			Gemini	0.192	0.373	0.491	0.558	0.429
		SSIM (↑)	GPT	0.189	0.354	0.050	**0.119**	**0.092**
			Gemini	0.232	0.403	0.266	0.339	0.266
		LPIPS (↓)	GPT	**0.173**	**0.346**	0.059	-0.024	-0.010
			Gemini	0.292	0.466	-0.153	-0.188	-0.143
		Avg	GPT	0.208	0.383	0.032	0.020	0.019
			Gemini	0.334	0.507	-0.083	-0.062	-0.049
	Spatial Relationship	L1 (↓)	GPT	0.253	0.440	**-0.130**	-0.144	-0.109
			Gemini	0.426	0.606	-0.269	-0.246	-0.181
		L2 (↓)	GPT	0.294	0.472	-0.122	**-0.167**	**-0.115**
			Gemini	0.460	0.620	-0.197	-0.192	-0.141
		PSNR (↑)	GPT	0.171	0.348	0.116	0.167	0.115
			Gemini	0.254	0.423	0.236	0.192	0.141
		SSIM (↑)	GPT	0.175	0.340	0.072	0.061	0.038
			Gemini	0.229	0.408	0.335	0.387	0.309
		LPIPS (↓)	GPT	**0.170**	**0.333**	0.015	0.053	0.033
			Gemini	0.354	0.548	-0.295	-0.271	-0.215
		Avg	GPT	0.213	0.387	-0.010	-0.006	-0.008
			Gemini	0.345	0.521	-0.038	-0.026	-0.017
	Texture and Detail	L1 (↓)	GPT	0.179	0.370	-0.177	**-0.287**	**-0.202**
			Gemini	0.258	0.445	-0.005	0.062	0.072
		L2 (↓)	GPT	0.217	0.413	**-0.244**	-0.286	-0.196
			Gemini	0.309	0.476	-0.140	-0.061	-0.021
		PSNR (↑)	GPT	**0.118**	**0.280**	0.147	0.286	0.196
			Gemini	0.237	0.402	0.041	0.061	0.021
		SSIM (↑)	GPT	0.121	0.285	0.097	0.173	0.109
			Gemini	0.271	0.463	-0.135	-0.045	-0.041
		LPIPS (↓)	GPT	0.124	0.285	-0.053	-0.168	-0.131
			Gemini	0.196	0.386	0.151	0.157	0.144
		Avg	GPT	0.152	0.327	-0.046	-0.056	-0.045
			Gemini	0.254	0.434	-0.018	0.035	0.035
	Image Quality	L1 (↓)	GPT	0.331	0.488	-0.150	-0.221	-0.184
			Gemini	0.323	0.505	0.160	0.092	0.087
		L2 (↓)	GPT	0.385	0.550	-0.087	-0.128	-0.104
			Gemini	0.392	0.560	0.057	0.043	0.042
		PSNR (↑)	GPT	0.222	0.387	0.070	0.128	0.104
			Gemini	0.326	0.506	-0.102	-0.043	-0.042
		SSIM (↑)	GPT	**0.182**	**0.354**	**0.355**	**0.310**	**0.246**
			Gemini	0.370	0.560	-0.312	-0.279	-0.215
		LPIPS (↓)	GPT	0.241	0.414	-0.217	-0.267	-0.215
			Gemini	0.243	0.428	0.204	0.201	0.164
		Avg	GPT	0.272	0.439	-0.006	-0.036	-0.031
			Gemini	0.331	0.512	0.001	0.003	0.007
	Color and Lighting	L1 (↓)	GPT	0.267	0.467	**-0.321**	**-0.338**	**-0.256**
			Gemini	0.285	0.453	-0.078	0.016	0.018
		L2 (↓)	GPT	0.306	0.508	-0.316	-0.256	-0.180
			Gemini	0.338	0.490	-0.193	-0.084	-0.060
		PSNR (↑)	GPT	0.146	0.313	0.221	0.256	0.180
			Gemini	0.228	0.385	0.079	0.084	0.060
		SSIM (↑)	GPT	**0.136**	**0.274**	0.267	0.254	0.207
			Gemini	0.209	0.388	0.151	0.123	0.081
		LPIPS (↓)	GPT	0.182	0.360	-0.144	-0.156	-0.109
			Gemini	0.211	0.401	0.083	0.061	0.039
		Avg	GPT	0.207	0.384	-0.059	-0.048	-0.032
			Gemini	0.254	0.423	0.008	0.040	0.028
	Seamlessness	L1 (↓)	GPT	0.227	0.430	-0.329	**-0.418**	**-0.291**
			Gemini	0.326	0.504	-0.114	-0.102	-0.066
		L2 (↓)	GPT	0.267	0.467	**-0.382**	-0.389	-0.280
			Gemini	0.381	0.540	-0.206	-0.150	-0.106
		PSNR (↑)	GPT	**0.111**	**0.268**	0.313	0.389	0.280
			Gemini	0.243	0.417	0.096	0.150	0.106
		SSIM (↑)	GPT	0.141	0.309	0.067	0.107	0.071
			Gemini	0.273	0.455	-0.051	-0.019	-0.009
		LPIPS (↓)	GPT	0.142	0.329	-0.049	-0.127	-0.086
			Gemini	0.255	0.450	-0.037	-0.037	-0.032
		Avg	GPT	0.178	0.361	-0.076	-0.088	-0.061
			Gemini	0.296	0.473	-0.062	-0.032	-0.021

Table 18: Results of Instruction Fidelity showing five evaluation metrics for all factors using traditional metrics (L1, L2, PSNR, SSIM, LPIPS) (OFFLINE setting, refer to Figure 8 for prompt). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

	Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
INSTRUCTION FIDELITY	Alignment	L1 (↓)	GPT	0.304	0.471	-0.037	-0.015	-0.008
			Gemini	0.173	0.329	-0.161	-0.107	-0.086
		L2 (↓)	GPT	0.361	0.517	-0.083	**-0.145**	**-0.098**
			Gemini	0.152	0.307	-0.173	-0.179	-0.135
		PSNR (↑)	GPT	**0.220**	**0.385**	**0.097**	0.145	0.098
			Gemini	0.206	0.375	0.149	0.179	0.135
		SSIM (↑)	GPT	0.251	0.419	-0.094	-0.053	-0.042
			Gemini	0.182	0.359	0.288	0.282	0.224
		LPIPS (↓)	GPT	0.223	0.389	-0.010	0.122	0.093
			Gemini	0.238	0.404	-0.277	-0.217	-0.158
		Avg	GPT	0.272	0.436	-0.025	0.011	0.009
			Gemini	0.190	0.355	-0.035	-0.008	-0.004
	Completeness	L1 (↓)	GPT	0.345	0.511	-0.078	-0.106	-0.075
			Gemini	0.332	0.504	-0.119	-0.010	-0.017
		L2 (↓)	GPT	0.414	0.562	**-0.168**	**-0.263**	**-0.212**
			Gemini	0.360	0.513	-0.179	-0.081	-0.076
		PSNR (↑)	GPT	**0.233**	**0.398**	0.158	0.263	0.212
			Gemini	0.276	0.436	0.129	0.081	0.076
		SSIM (↑)	GPT	0.272	0.455	-0.068	-0.041	-0.045
			Gemini	0.247	0.418	0.224	0.246	0.215
		LPIPS (↓)	GPT	0.262	0.445	-0.088	0.014	0.015
			Gemini	0.386	0.575	-0.415	-0.393	-0.314
		Avg	GPT	0.305	0.474	-0.049	-0.027	-0.021
			Gemini	0.320	0.489	-0.072	-0.031	-0.023
	Plausibility	L1 (↓)	GPT	0.273	0.448	0.000	-0.046	-0.045
			Gemini	0.361	0.541	-0.266	-0.222	-0.160
		L2 (↓)	GPT	0.326	0.492	0.001	-0.063	-0.040
			Gemini	0.413	0.585	-0.344	-0.270	-0.206
		PSNR (↑)	GPT	0.212	0.403	-0.013	0.063	0.040
			Gemini	0.209	0.389	0.259	0.270	0.206
		SSIM (↑)	GPT	0.201	0.370	**0.034**	**0.070**	**0.062**
			Gemini	0.220	0.409	0.196	0.188	0.148
		LPIPS (↓)	GPT	**0.193**	**0.355**	0.003	0.027	0.029
			Gemini	0.262	0.451	-0.062	-0.077	-0.057
		Avg	GPT	0.241	0.414	0.005	0.010	0.009
			Gemini	0.293	0.475	-0.043	-0.022	-0.014

Table 19: Results of Image Preservation showing five evaluation metrics for all factors using traditional metrics (L1, L2, PSNR, SSIM, LPIPS) (ONLINE setting, refer to Figure 8 for prompt)). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

	Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
IMAGE PRESERVATION	Unchanged Regions	L1 (↓)	GPT	0.345	0.526	**0.181**	**0.164**	**0.136**
			Gemini	0.343	0.510	0.075	-0.002	0.004
		L2 (↓)	GPT	0.408	0.575	0.127	0.106	0.088
			Gemini	0.389	0.551	0.122	0.100	0.076
		PSNR (↑)	GPT	0.369	0.550	-0.116	-0.106	-0.088
			Gemini	0.346	0.521	-0.142	-0.100	-0.076
		SSIM (↑)	GPT	0.342	0.523	-0.072	-0.067	-0.056
			Gemini	0.261	0.445	0.214	0.260	0.205
		LPIPS (↓)	GPT	**0.269**	**0.437**	0.030	0.029	0.024
			Gemini	0.290	0.475	-0.219	-0.272	-0.220
		Avg	GPT	0.347	0.522	0.030	0.025	0.021
			Gemini	0.326	0.500	0.010	-0.003	-0.002
	Global Consistency	L1 (↓)	GPT	0.395	0.576	-0.080	-0.029	-0.024
			Gemini	0.360	0.555	-0.155	-0.162	-0.116
		L2 (↓)	GPT	0.456	0.624	-0.130	-0.029	-0.024
			Gemini	0.409	0.590	-0.094	-0.047	-0.041
		PSNR (↑)	GPT	0.354	0.535	-0.043	0.029	0.024
			Gemini	0.298	0.488	0.006	0.047	0.041
		SSIM (↑)	GPT	0.337	0.519	-0.049	-0.077	-0.064
			Gemini	0.263	0.440	0.100	0.157	0.123
		LPIPS (↓)	GPT	**0.316**	**0.484**	**-0.187**	**-0.193**	**-0.160**
			Gemini	0.235	0.409	-0.116	-0.234	-0.177
		Avg	GPT	0.372	0.548	-0.098	-0.060	-0.050
			Gemini	0.313	0.496	-0.052	-0.048	-0.034
	Identity Preservation	L1 (↓)	GPT	0.217	0.399	-0.048	-0.024	-0.020
			Gemini	0.366	0.542	0.040	-0.123	-0.109
		L2 (↓)	GPT	**0.167**	**0.335**	-0.131	-0.053	-0.042
			Gemini	0.426	0.582	0.081	-0.011	-0.008
		PSNR (↑)	GPT	0.276	0.457	0.026	0.023	0.018
			Gemini	0.339	0.511	-0.112	0.011	0.008
		SSIM (↑)	GPT	0.305	0.486	0.065	0.061	0.048
			Gemini	0.274	0.453	0.205	0.332	0.259
		LPIPS (↓)	GPT	0.427	0.595	**-0.156**	**-0.115**	**-0.091**
			Gemini	0.255	0.432	-0.211	-0.312	-0.259
		Avg	GPT	0.278	0.454	-0.049	-0.022	-0.017
			Gemini	0.332	0.504	0.001	-0.021	-0.022

Table 20: Results of Edit Quality showing five evaluation metrics for all factors using traditional metrics (L1, L2, PSNR, SSIM, LPIPS) (ONLINE setting, refer to Figure 8 for prompt)). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
Overall	L1 (↓)	GPT	0.309	0.498	0.195	0.105	0.078
		Gemini	0.388	0.571	-0.200	-0.232	-0.184
	L2 (↓)	GPT	0.366	0.536	0.221	**0.177**	**0.130**
		Gemini	0.443	0.615	-0.194	-0.196	-0.169
	PSNR (↑)	GPT	0.338	0.525	**-0.322**	-0.177	-0.130
		Gemini	0.296	0.480	0.132	0.204	0.177
	SSIM (↑)	GPT	0.271	0.462	0.010	0.009	0.026
		Gemini	0.256	0.450	0.245	0.316	0.256
	LPIPS (↓)	GPT	**0.202**	**0.374**	0.051	-0.108	-0.085
		Gemini	0.256	0.433	-0.119	-0.238	-0.177
	Avg	GPT	0.297	0.479	0.031	0.001	0.004
		Gemini	0.328	0.510	-0.027	-0.029	-0.019
Spatial Relationship	L1 (↓)	GPT	0.340	0.517	0.057	-0.109	-0.096
		Gemini	0.405	0.588	-0.111	-0.065	-0.052
	L2 (↓)	GPT	0.405	0.576	0.038	-0.059	-0.053
		Gemini	0.465	0.636	-0.058	-0.003	0.000
	PSNR (↑)	GPT	0.323	0.516	-0.196	0.059	0.053
		Gemini	0.339	0.523	0.031	0.016	0.010
	SSIM (↑)	GPT	0.272	0.443	0.051	0.113	0.089
		Gemini	0.276	0.454	0.319	0.314	0.251
	LPIPS (↓)	GPT	**0.237**	**0.405**	**-0.254**	**-0.349**	**-0.287**
		Gemini	0.268	0.435	-0.194	-0.195	-0.167
	Avg	GPT	0.315	0.491	-0.061	-0.069	-0.059
		Gemini	0.351	0.527	-0.003	0.013	0.008
Texture and Detail	L1 (↓)	GPT	0.274	0.458	0.020	0.008	0.011
		Gemini	0.378	0.553	-0.395	-0.354	-0.282
	L2 (↓)	GPT	0.323	0.503	-0.018	0.001	0.000
		Gemini	0.431	0.600	-0.375	-0.208	-0.168
	PSNR (↑)	GPT	0.268	0.444	-0.102	-0.001	0.000
		Gemini	0.255	0.433	0.218	0.216	0.174
	SSIM (↑)	GPT	**0.230**	0.420	0.038	0.051	0.051
		Gemini	0.251	0.438	0.104	0.171	0.108
	LPIPS (↓)	GPT	0.237	**0.402**	**-0.214**	**-0.215**	**-0.158**
		Gemini	0.226	0.383	-0.120	-0.077	-0.072
	Avg	GPT	0.266	0.445	-0.055	-0.031	-0.019
		Gemini	0.308	0.481	-0.114	-0.050	-0.048
Image Quality	L1 (↓)	GPT	0.387	0.568	-0.004	0.067	0.056
		Gemini	0.403	0.584	0.031	-0.015	-0.013
	L2 (↓)	GPT	0.452	0.619	-0.052	0.057	0.047
		Gemini	0.467	0.634	0.077	0.093	0.077
	PSNR (↑)	GPT	0.359	0.540	-0.056	-0.057	-0.047
		Gemini	0.374	0.555	-0.132	-0.093	-0.077
	SSIM (↑)	GPT	0.362	0.544	**-0.177**	**-0.202**	**-0.167**
		Gemini	0.314	0.495	0.129	0.124	0.103
	LPIPS (↓)	GPT	**0.273**	**0.441**	-0.019	-0.036	-0.030
		Gemini	0.248	0.416	0.016	0.000	0.000
	Avg	GPT	0.367	0.542	-0.062	-0.034	-0.028
		Gemini	0.361	0.537	0.024	0.022	0.018
Color and Lighting	L1 (↓)	GPT	0.297	0.476	-0.092	-0.106	-0.084
		Gemini	0.358	0.538	-0.089	-0.055	-0.038
	L2 (↓)	GPT	0.351	0.525	-0.137	-0.108	-0.090
		Gemini	0.412	0.583	-0.119	-0.016	-0.010
	PSNR (↑)	GPT	0.252	0.434	0.006	0.102	0.084
		Gemini	0.324	0.509	-0.078	0.016	0.010
	SSIM (↑)	GPT	**0.222**	**0.387**	0.097	0.119	0.101
		Gemini	0.284	0.454	0.043	0.026	0.017
	LPIPS (↓)	GPT	0.236	0.410	**-0.263**	**-0.314**	**-0.246**
		Gemini	0.274	0.446	-0.151	-0.151	-0.127
	Avg	GPT	0.272	0.446	-0.078	-0.061	-0.047
		Gemini	0.330	0.506	-0.079	-0.036	-0.030
Seamlessness	L1 (↓)	GPT	0.313	0.501	-0.058	-0.077	-0.056
		Gemini	0.368	0.541	-0.189	-0.145	-0.116
	L2 (↓)	GPT	0.362	0.538	-0.072	-0.052	-0.050
		Gemini	0.419	0.589	-0.208	-0.096	-0.078
	PSNR (↑)	GPT	0.279	0.453	-0.027	0.046	0.044
		Gemini	0.300	0.481	0.011	0.103	0.084
	SSIM (↑)	GPT	**0.250**	0.431	0.050	0.050	0.050
		Gemini	0.274	0.460	0.059	0.019	0.006
	LPIPS (↓)	GPT	0.251	**0.427**	**-0.180**	**-0.204**	**-0.157**
		Gemini	0.262	0.422	-0.115	-0.072	-0.071
	Avg	GPT	0.291	0.470	-0.057	-0.047	-0.034
		Gemini	0.325	0.499	-0.088	-0.038	-0.035

Table 21: Results of Instruction Fidelity showing five evaluation metrics for all factors using traditional metrics (L1, L2, PSNR, SSIM, LPIPS) (ONLINE setting, refer to Figure 8 for prompt)). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

	Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
INSTRUCTION FIDELITY	Alignment	L1 (↓)	GPT	0.378	0.547	-0.004	-0.061	-0.054
			Gemini	0.403	0.591	-0.019	0.018	0.019
		L2 (↓)	GPT	0.440	0.593	0.013	-0.104	-0.071
			Gemini	0.471	0.646	0.006	0.067	0.057
		PSNR (↑)	GPT	0.331	0.513	0.007	0.104	0.071
			Gemini	0.353	0.539	-0.030	-0.067	-0.057
		SSIM (↑)	GPT	0.320	0.508	-0.083	0.073	0.071
			Gemini	0.312	0.494	0.103	0.041	0.032
		LPIPS (↓)	GPT	**0.184**	**0.343**	**0.359**	**0.140**	**0.112**
			Gemini	0.229	0.391	0.099	0.149	0.121
		Avg	GPT	0.331	0.501	0.058	0.030	0.026
			Gemini	0.354	0.532	0.032	0.042	0.034
	Completeness	L1 (↓)	GPT	0.378	0.547	-0.004	-0.061	-0.054
			Gemini	0.430	0.611	-0.194	-0.247	-0.205
		L2 (↓)	GPT	0.440	0.593	0.013	-0.104	-0.071
			Gemini	0.499	0.666	-0.142	-0.161	-0.134
		PSNR (↑)	GPT	0.331	0.513	0.007	0.104	0.071
			Gemini	0.345	0.526	0.152	0.161	0.134
		SSIM (↑)	GPT	0.320	0.508	-0.083	0.073	0.071
			Gemini	0.311	0.492	0.220	0.225	0.187
		LPIPS (↓)	GPT	**0.184**	**0.343**	**0.359**	**0.140**	**0.112**
			Gemini	0.243	0.411	0.006	-0.011	-0.009
		Avg	GPT	0.331	0.501	0.058	0.030	0.026
			Gemini	0.366	0.541	0.008	-0.007	-0.005
	Plausibility	L1 (↓)	GPT	0.311	0.495	0.140	0.010	-0.016
			Gemini	0.381	0.575	-0.213	-0.183	-0.130
		L2 (↓)	GPT	0.365	0.533	0.152	**0.074**	**0.043**
			Gemini	0.417	0.598	-0.192	-0.104	-0.067
		PSNR (↑)	GPT	0.338	0.525	**-0.269**	-0.074	-0.043
			Gemini	0.300	0.482	0.060	0.113	0.073
		SSIM (↑)	GPT	0.280	0.452	-0.026	-0.005	-0.003
			Gemini	0.230	0.410	0.314	0.299	0.244
		LPIPS (↓)	GPT	**0.202**	**0.374**	0.102	-0.024	-0.023
			Gemini	0.321	0.492	-0.228	-0.246	-0.181
		Avg	GPT	0.299	0.476	-0.000	-0.004	-0.008
			Gemini	0.330	0.511	-0.052	-0.024	-0.012

Table 22: Results of Image Preservation showing evaluation metrics using mask-based metrics (Mask SSIM, Mask LPIPS, Background Consistency) (OFFLINE setting, refer to Figure 8 for prompt)). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
Unchanged Regions	Mask SSIM (↑)	GPT	**0.175**	**0.338**	**0.264**	**0.180**	**0.138**
		Gemini	0.276	0.442	-0.004	-0.022	-0.009
	Mask LPIPS (↓)	GPT	0.199	0.385	0.036	0.104	0.073
		Gemini	0.253	0.454	0.042	-0.044	-0.027
	BG Consistency (↑)	GPT	**0.175**	**0.338**	**0.264**	**0.180**	**0.138**
		Gemini	0.276	0.442	-0.004	-0.022	-0.009
	Avg	GPT	0.183	0.354	0.188	0.155	0.116
		Gemini	0.268	0.446	0.011	-0.029	-0.015
Global Consistency	Mask SSIM (↑)	GPT	**0.178**	**0.340**	**0.091**	**0.084**	**0.060**
		Gemini	0.298	0.484	-0.062	-0.041	-0.032
	Mask LPIPS (↓)	GPT	0.182	0.359	-0.020	-0.044	-0.055
		Gemini	0.272	0.473	-0.011	-0.015	-0.009
	BG Consistency (↑)	GPT	**0.178**	**0.340**	**0.091**	**0.084**	**0.060**
		Gemini	0.298	0.484	-0.062	-0.041	-0.032
	Avg	GPT	0.179	0.346	0.054	0.041	0.022
		Gemini	0.289	0.480	-0.045	-0.032	-0.024
Identity Preservation	Mask SSIM (↑)	GPT	**0.122**	**0.283**	**0.529**	**0.533**	**0.423**
		Gemini	0.292	0.467	-0.034	0.017	0.013
	Mask LPIPS (↓)	GPT	0.248	0.428	-0.266	-0.248	-0.165
		Gemini	0.264	0.452	0.041	-0.002	0.019
	BG Consistency (↑)	GPT	**0.122**	**0.283**	**0.529**	**0.533**	**0.423**
		Gemini	0.292	0.467	-0.034	0.017	0.013
	Avg	GPT	0.164	0.331	0.264	0.273	0.227
		Gemini	0.283	0.462	-0.009	0.011	0.015

Table 23: Results of Edit Quality showing evaluation metrics using mask-based metrics (Mask SSIM, Mask LPIPS, Background Consistency) (OFFLINE setting, refer to Figure 8 for prompt)). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
Scale Realism	Mask SSIM (↑)	GPT	0.189	0.354	0.050	**0.119**	**0.092**
		Gemini	0.232	0.403	0.266	0.339	0.266
	Mask LPIPS (↓)	GPT	**0.173**	**0.346**	**0.059**	-0.024	-0.010
		Gemini	0.292	0.466	-0.153	-0.188	-0.143
	BG Consistency (↑)	GPT	0.189	0.354	0.050	**0.119**	**0.092**
		Gemini	0.232	0.403	0.266	0.339	0.266
	Avg	GPT	0.184	0.351	0.053	0.071	0.058
		Gemini	0.252	0.424	0.126	0.163	0.130
Spatial Relationship	Mask SSIM (↑)	GPT	0.175	0.340	**0.072**	**0.061**	**0.038**
		Gemini	0.229	0.408	0.335	0.387	0.309
	Mask LPIPS (↓)	GPT	**0.170**	**0.333**	0.015	0.053	0.033
		Gemini	0.354	0.548	-0.295	-0.271	-0.215
	BG Consistency (↑)	GPT	0.175	0.340	**0.072**	**0.061**	**0.038**
		Gemini	0.229	0.408	0.335	0.387	0.309
	Avg	GPT	0.173	0.338	0.053	0.058	0.036
		Gemini	0.271	0.455	0.125	0.168	0.134
Texture and Detail	Mask SSIM (↑)	GPT	**0.121**	**0.285**	**0.097**	**0.173**	0.109
		Gemini	0.271	0.463	-0.135	-0.045	-0.041
	Mask LPIPS (↓)	GPT	0.124	0.285	-0.053	-0.168	**-0.131**
		Gemini	0.196	0.386	0.151	0.157	0.144
	BG Consistency (↑)	GPT	**0.121**	**0.285**	**0.097**	**0.173**	0.109
		Gemini	0.271	0.463	-0.135	-0.045	-0.041
	Avg	GPT	0.122	0.285	0.047	0.059	0.029
		Gemini	0.246	0.437	-0.040	0.022	0.021
Image Quality	Mask SSIM (↑)	GPT	**0.182**	**0.354**	**0.355**	**0.310**	**0.246**
		Gemini	0.370	0.560	-0.312	-0.279	-0.215
	Mask LPIPS (↓)	GPT	0.241	0.414	-0.217	-0.267	-0.215
		Gemini	0.243	0.428	0.204	0.201	0.164
	BG Consistency (↑)	GPT	**0.182**	**0.354**	**0.355**	**0.310**	**0.246**
		Gemini	0.370	0.560	-0.312	-0.279	-0.215
	Avg	GPT	0.202	0.374	0.164	0.118	0.092
		Gemini	0.328	0.516	-0.140	-0.119	-0.089
Color and Lighting	Mask SSIM (↑)	GPT	**0.136**	**0.274**	**0.267**	**0.254**	**0.207**
		Gemini	0.209	0.388	0.151	0.123	0.081
	Mask LPIPS (↓)	GPT	0.182	0.360	-0.144	-0.156	-0.109
		Gemini	0.211	0.401	0.083	0.061	0.039
	BG Consistency (↑)	GPT	**0.136**	**0.274**	**0.267**	**0.254**	**0.207**
		Gemini	0.209	0.388	0.151	0.123	0.081
	Avg	GPT	0.151	0.303	0.130	0.117	0.102
		Gemini	0.210	0.392	0.128	0.102	0.067
Seamlessness	Mask SSIM (↑)	GPT	**0.141**	**0.309**	**0.067**	0.107	0.071
		Gemini	0.273	0.455	-0.051	-0.019	-0.009
	Mask LPIPS (↓)	GPT	0.142	0.329	-0.049	**-0.127**	**-0.086**
		Gemini	0.255	0.450	-0.037	-0.037	-0.032
	BG Consistency (↑)	GPT	**0.141**	**0.309**	**0.067**	0.107	0.071
		Gemini	0.273	0.455	-0.051	-0.019	-0.009
	Avg	GPT	0.141	0.316	0.028	0.029	0.019
		Gemini	0.267	0.453	-0.046	-0.025	-0.017

Table 24: Results of Instruction Fidelity showing evaluation metrics using mask-based metrics (Mask SSIM, Mask LPIPS, Background Consistency) (OFFLINE setting, refer to Figure 8 for prompt)). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
Alignment	Mask SSIM (↑)	GPT	0.251	0.419	**-0.094**	-0.053	-0.042
		Gemini	0.182	0.359	0.288	0.282	0.224
	Mask LPIPS (↓)	GPT	**0.223**	**0.389**	-0.010	**0.122**	**0.093**
		Gemini	0.238	0.404	-0.277	-0.217	-0.158
	BG Consistency (↑)	GPT	0.251	0.419	**-0.094**	-0.053	-0.042
		Gemini	0.182	0.359	0.288	0.282	0.224
	Avg	GPT	0.242	0.409	-0.066	0.005	0.003
		Gemini	0.201	0.374	0.100	0.116	0.097
Completeness	Mask SSIM (↑)	GPT	0.272	0.455	-0.068	**-0.041**	**-0.045**
		Gemini	0.247	0.418	0.224	0.246	0.215
	Mask LPIPS (↓)	GPT	**0.262**	**0.445**	**-0.088**	0.014	0.015
		Gemini	0.386	0.575	-0.415	-0.393	-0.314
	BG Consistency (↑)	GPT	0.272	0.455	-0.068	**-0.041**	**-0.045**
		Gemini	0.247	0.418	0.224	0.246	0.215
	Avg	GPT	0.269	0.452	-0.075	-0.023	-0.025
		Gemini	0.293	0.470	0.011	0.033	0.039
Plausibility	Mask SSIM (↑)	GPT	0.201	0.370	**0.034**	**0.070**	**0.062**
		Gemini	0.220	0.409	0.196	0.188	0.148
	Mask LPIPS (↓)	GPT	**0.193**	**0.355**	0.003	0.027	0.029
		Gemini	0.262	0.451	-0.062	-0.077	-0.057
	BG Consistency (↑)	GPT	0.201	0.370	**0.034**	**0.070**	**0.062**
		Gemini	0.220	0.409	0.196	0.188	0.148
	Avg	GPT	0.198	0.365	0.024	0.056	0.051
		Gemini	0.234	0.423	0.110	0.100	0.080

Table 25: Results of Image Preservation using mask-based metrics (Mask SSIM, Mask LPIPS, Background Consistency) (ONLINE setting, refer to Figure 8 for prompt)). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

	Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
Image Preservation	Unchanged Regions	Mask SSIM (↑)	GPT	0.342	0.523	**-0.072**	**-0.067**	**-0.056**
			Gemini	0.261	0.445	0.214	0.260	0.205
		Mask LPIPS (↓)	GPT	**0.269**	**0.437**	0.030	0.029	0.024
			Gemini	0.290	0.475	-0.219	-0.272	-0.220
		BG Consistency (↑)	GPT	0.342	0.523	**-0.072**	**-0.067**	**-0.056**
			Gemini	0.261	0.445	0.214	0.260	0.205
		Avg	GPT	0.318	0.494	-0.038	-0.035	-0.029
			Gemini	0.271	0.455	0.070	0.083	0.063
	Global Consistency	Mask SSIM (↑)	GPT	0.337	0.519	-0.049	-0.077	-0.064
			Gemini	0.263	0.440	0.100	0.157	0.123
		Mask LPIPS (↓)	GPT	**0.316**	**0.484**	**-0.187**	**-0.193**	**-0.160**
			Gemini	0.235	0.409	-0.116	-0.234	-0.177
		BG Consistency (↑)	GPT	0.337	0.519	-0.049	-0.077	-0.064
			Gemini	0.263	0.440	0.100	0.157	0.123
		Avg	GPT	0.330	0.507	-0.095	-0.116	-0.096
			Gemini	0.254	0.430	0.028	0.027	0.023
	Identity Preservation	Mask SSIM (↑)	GPT	**0.305**	**0.486**	0.065	0.061	0.048
			Gemini	0.274	0.453	0.205	0.332	0.259
		Mask LPIPS (↓)	GPT	0.427	0.595	**-0.156**	**-0.115**	**-0.091**
			Gemini	0.255	0.432	-0.211	-0.312	-0.259
		BG Consistency (↑)	GPT	**0.305**	**0.486**	0.065	0.061	0.048
			Gemini	0.274	0.453	0.205	0.332	0.259
		Avg	GPT	0.346	0.522	-0.009	0.002	0.002
			Gemini	0.268	0.446	0.066	0.117	0.086

Table 26: Results of Edit Quality using mask-based metrics (Mask SSIM, Mask LPIPS, Background Consistency) (ONLINE setting, refer to Figure 8 for prompt)). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

	Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
EDIT QUALITY	Scale Realism	Mask SSIM (↑)	GPT	0.271	0.462	0.010	0.009	0.026
			Gemini	0.256	0.450	0.245	0.316	0.256
		Mask LPIPS (↓)	GPT	**0.202**	**0.374**	**0.051**	**-0.108**	**-0.085**
			Gemini	0.256	0.433	-0.119	-0.238	-0.177
		BG Consistency (↑)	GPT	0.271	0.462	0.010	0.009	0.026
			Gemini	0.256	0.450	0.245	0.316	0.256
		Avg	GPT	0.248	0.433	0.024	-0.030	-0.011
			Gemini	0.256	0.444	0.124	0.131	0.112
	Spatial Relationship	Mask SSIM (↑)	GPT	0.272	0.443	0.051	0.113	0.089
			Gemini	0.276	0.454	0.319	0.314	0.251
		Mask LPIPS (↓)	GPT	**0.237**	**0.405**	**-0.254**	**-0.349**	**-0.287**
			Gemini	0.268	0.435	-0.194	-0.195	-0.167
		BG Consistency (↑)	GPT	0.272	0.443	0.051	0.113	0.089
			Gemini	0.276	0.454	0.319	0.314	0.251
		Avg	GPT	0.260	0.430	-0.051	-0.041	-0.036
			Gemini	0.273	0.448	0.148	0.144	0.112
	Texture and Detail	Mask SSIM (↑)	GPT	**0.230**	0.420	0.038	0.051	0.051
			Gemini	0.251	0.438	0.104	0.171	0.108
		Mask LPIPS (↓)	GPT	0.237	**0.402**	**-0.214**	**-0.215**	**-0.158**
			Gemini	0.226	0.383	-0.120	-0.077	-0.072
		BG Consistency (↑)	GPT	**0.230**	0.420	0.038	0.051	0.051
			Gemini	0.251	0.438	0.104	0.171	0.108
		Avg	GPT	0.232	0.414	-0.046	-0.038	-0.019
			Gemini	0.243	0.420	0.029	0.088	0.048
	Image Quality	Mask SSIM (↑)	GPT	0.362	0.544	**-0.177**	**-0.202**	**-0.167**
			Gemini	0.314	0.495	0.129	0.124	0.103
		Mask LPIPS (↓)	GPT	**0.273**	**0.441**	-0.019	-0.036	-0.030
			Gemini	0.248	0.416	0.016	0.000	0.000
		BG Consistency (↑)	GPT	0.362	0.544	**-0.177**	**-0.202**	**-0.167**
			Gemini	0.314	0.495	0.129	0.124	0.103
		Avg	GPT	0.332	0.510	-0.124	-0.147	-0.121
			Gemini	0.292	0.469	0.091	0.083	0.069
	Color and Lighting	Mask SSIM (↑)	GPT	**0.222**	**0.387**	0.097	0.119	0.101
			Gemini	0.284	0.454	0.043	0.026	0.017
		Mask LPIPS (↓)	GPT	0.236	0.410	**-0.263**	**-0.314**	**-0.246**
			Gemini	0.274	0.446	-0.151	-0.151	-0.127
		BG Consistency (↑)	GPT	**0.222**	**0.387**	0.097	0.119	0.101
			Gemini	0.284	0.454	0.043	0.026	0.017
		Avg	GPT	0.227	0.395	-0.023	-0.025	-0.015
			Gemini	0.281	0.451	-0.022	-0.033	-0.031
	Seamlessness	Mask SSIM (↑)	GPT	**0.250**	0.431	0.050	0.050	0.050
			Gemini	0.274	0.460	0.059	0.019	0.006
		Mask LPIPS (↓)	GPT	0.251	**0.427**	**-0.180**	**-0.204**	**-0.157**
			Gemini	0.262	0.422	-0.115	-0.072	-0.071
		BG Consistency (↑)	GPT	**0.250**	0.431	0.050	0.050	0.050
			Gemini	0.274	0.460	0.059	0.019	0.006
		Avg	GPT	0.250	0.430	-0.027	-0.035	-0.019
			Gemini	0.270	0.447	0.001	-0.011	-0.020

Table 27: Results of Instruction Fidelity using mask-based metrics (Mask SSIM, Mask LPIPS, Background Consistency) (ONLINE setting, refer to Figure 8 for prompt)). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

	Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
INSTRUCTION FIDELITY	Alignment	Mask SSIM (↑)	GPT	0.320	0.508	-0.083	0.073	0.071
			Gemini	0.312	0.494	0.103	0.041	0.032
		Mask LPIPS (↓)	GPT	**0.184**	**0.343**	**0.359**	**0.140**	**0.112**
			Gemini	0.229	0.391	0.099	0.149	0.121
		BG Consistency (↑)	GPT	0.320	0.508	-0.083	0.073	0.071
			Gemini	0.312	0.494	0.103	0.041	0.032
		Avg	GPT	0.275	0.453	0.064	0.095	0.085
			Gemini	0.284	0.460	0.102	0.077	0.062
	Completeness	Mask SSIM (↑)	GPT	0.320	0.508	-0.083	0.073	0.071
			Gemini	0.311	0.492	0.220	0.225	0.187
		Mask LPIPS (↓)	GPT	**0.184**	**0.343**	**0.359**	**0.140**	**0.112**
			Gemini	0.243	0.411	0.006	-0.011	-0.009
		BG Consistency (↑)	GPT	0.320	0.508	-0.083	0.073	0.071
			Gemini	0.311	0.492	0.220	0.225	0.187
		Avg	GPT	0.275	0.453	0.064	0.095	0.085
			Gemini	0.288	0.465	0.149	0.146	0.122
	Plausibility	Mask SSIM (↑)	GPT	0.280	0.452	-0.026	-0.005	-0.003
			Gemini	0.230	0.410	0.314	0.299	0.244
		Mask LPIPS (↓)	GPT	**0.202**	**0.374**	**0.102**	**-0.024**	**-0.023**
			Gemini	0.321	0.492	-0.228	-0.246	-0.181
		BG Consistency (↑)	GPT	0.280	0.452	-0.026	-0.005	-0.003
			Gemini	0.230	0.410	0.314	0.299	0.244
		Avg	GPT	0.254	0.426	0.017	-0.011	-0.010
			Gemini	0.260	0.437	0.133	0.117	0.102

Table 28: Results of Image Preservation using semantic similarity metrics (CLIP Text, CLIP Image, DINO Image) (OFFLINE setting, refer to Figure 8 for prompt)). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

	Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
IMAGE PRESERVATION	Unchanged Regions	CLIP Text (↑)	GPT	**0.102**	0.292	**0.241**	**0.224**	**0.171**
			Gemini	0.172	0.398	0.245	0.203	0.144
		CLIP Image (↑)	GPT	0.109	**0.230**	-0.121	-0.070	-0.051
			Gemini	0.219	0.335	-0.161	-0.080	-0.085
		DINO Image (↑)	GPT	0.128	0.302	0.130	-0.065	-0.062
			Gemini	0.240	0.403	-0.062	-0.078	-0.062
		Avg	GPT	0.113	0.275	0.083	0.030	0.019
			Gemini	0.210	0.379	0.007	0.015	-0.001
	Global Consistency	CLIP Text (↑)	GPT	**0.095**	**0.275**	0.160	0.135	0.097
			Gemini	0.181	0.411	0.111	0.143	0.103
		CLIP Image (↑)	GPT	0.147	0.295	**-0.230**	**-0.175**	**-0.129**
			Gemini	0.219	0.329	-0.065	-0.019	-0.021
		DINO Image (↑)	GPT	0.157	0.337	-0.051	-0.083	-0.071
			Gemini	0.230	0.397	0.016	-0.006	-0.003
		Avg	GPT	0.133	0.302	-0.040	-0.041	-0.034
			Gemini	0.210	0.379	0.021	0.039	0.026
	Identity Preservation	CLIP Text (↑)	GPT	0.118	0.318	-0.004	0.002	0.005
			Gemini	0.167	0.386	0.355	0.355	0.288
		CLIP Image (↑)	GPT	0.154	0.305	-0.070	-0.002	-0.005
			Gemini	0.179	0.285	-0.209	-0.113	-0.100
		DINO Image (↑)	GPT	**0.116**	**0.270**	**0.331**	**0.259**	**0.181**
			Gemini	0.246	0.406	-0.178	-0.113	-0.081
		Avg	GPT	0.129	0.298	0.086	0.086	0.060
			Gemini	0.197	0.359	-0.011	0.043	0.036

Table 29: Results of Edit Quality using semantic similarity metrics (CLIP Text, CLIP Image, DINO Image) (OFFLINE setting, refer to Figure 8 for prompt)). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

	Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
EDIT QUALITY	Scale Realism	CLIP Text (↑)	GPT	**0.114**	0.293	0.180	0.137	0.118
			Gemini	0.157	0.377	-0.273	-0.146	-0.123
		CLIP Image (↑)	GPT	0.207	0.361	-0.156	-0.120	-0.082
			Gemini	0.107	0.220	0.282	0.396	0.334
		DINO Image (↑)	GPT	0.128	**0.278**	**0.326**	**0.321**	**0.250**
			Gemini	0.111	0.255	0.463	0.455	0.375
		Avg	GPT	0.150	0.311	0.117	0.113	0.095
			Gemini	0.125	0.284	0.157	0.235	0.195
	Spatial Relationship	CLIP Text (↑)	GPT	**0.087**	**0.260**	0.037	0.150	0.089
			Gemini	0.195	0.428	0.130	0.165	0.128
		CLIP Image (↑)	GPT	0.132	0.272	-0.175	-0.121	-0.089
			Gemini	0.212	0.311	-0.145	-0.132	-0.107
		DINO Image (↑)	GPT	0.103	0.260	**0.269**	**0.199**	**0.140**
			Gemini	0.190	0.342	0.252	0.165	0.128
		Avg	GPT	0.107	0.264	0.044	0.076	0.047
			Gemini	0.199	0.360	0.079	0.066	0.050
	Texture and Detail	CLIP Text (↑)	GPT	**0.051**	**0.156**	0.036	0.035	0.027
			Gemini	0.174	0.377	-0.325	-0.294	-0.215
		CLIP Image (↑)	GPT	0.139	0.323	-0.092	-0.057	-0.049
			Gemini	0.279	0.409	0.036	0.208	0.144
		DINO Image (↑)	GPT	0.092	0.271	**0.224**	**0.242**	**0.180**
			Gemini	0.212	0.376	0.145	0.147	0.103
		Avg	GPT	0.094	0.250	0.056	0.073	0.053
			Gemini	0.222	0.387	-0.048	0.020	0.011
	Image Quality	CLIP Text (↑)	GPT	0.100	0.268	0.234	0.220	0.178
			Gemini	0.195	0.424	-0.175	-0.173	-0.138
		CLIP Image (↑)	GPT	**0.065**	**0.194**	-0.022	0.029	0.025
			Gemini	0.208	0.313	0.079	0.210	0.164
		DINO Image (↑)	GPT	0.104	0.274	**0.246**	**0.277**	**0.221**
			Gemini	0.243	0.414	-0.022	-0.034	-0.029
		Avg	GPT	0.090	0.245	0.153	0.175	0.141
			Gemini	0.215	0.384	-0.039	0.001	-0.001
	Color and Lighting	CLIP Text (↑)	GPT	**0.080**	**0.250**	-0.002	-0.090	-0.054
			Gemini	0.165	0.378	-0.424	-0.369	-0.260
		CLIP Image (↑)	GPT	0.129	0.277	**-0.185**	**-0.199**	**-0.136**
			Gemini	0.240	0.381	0.037	0.163	0.139
		DINO Image (↑)	GPT	0.127	0.310	0.051	0.044	0.038
			Gemini	0.228	0.385	-0.026	-0.070	-0.039
		Avg	GPT	0.112	0.279	-0.045	-0.082	-0.051
			Gemini	0.211	0.381	-0.138	-0.092	-0.053
	Seamlessness	CLIP Text (↑)	GPT	**0.061**	**0.204**	0.177	**0.221**	0.160
			Gemini	0.171	0.397	-0.113	-0.163	-0.112
		CLIP Image (↑)	GPT	0.131	0.289	-0.056	-0.051	-0.034
			Gemini	0.224	0.361	0.012	0.134	0.100
		DINO Image (↑)	GPT	0.098	0.277	**0.200**	0.208	**0.170**
			Gemini	0.231	0.402	-0.028	0.002	0.003
		Avg	GPT	0.097	0.257	0.107	0.126	0.099
			Gemini	0.209	0.387	-0.043	-0.009	-0.003

Table 30: Results of Instruction Fidelity using semantic similarity metrics (CLIP Text, CLIP Image, DINO Image) (OFFLINE setting, refer to Figure 8 for prompt)). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

	Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
INSTRUCTION FIDELITY	Alignment	CLIP Text (↑)	GPT	0.127	0.319	-0.133	-0.000	0.014
			Gemini	0.218	0.412	-0.247	-0.190	-0.141
		CLIP Image (↑)	GPT	**0.117**	**0.253**	**0.561**	**0.456**	**0.368**
			Gemini	0.504	0.670	0.220	0.142	0.130
		DINO Image (↑)	GPT	0.130	0.288	0.245	0.181	0.149
			Gemini	0.298	0.479	0.311	0.364	0.279
		Avg	GPT	0.125	0.287	0.224	0.212	0.177
			Gemini	0.340	0.520	0.095	0.105	0.089
	Completeness	CLIP Text (↑)	GPT	0.144	0.361	-0.075	0.015	0.015
			Gemini	0.252	0.490	-0.273	-0.227	-0.169
		CLIP Image (↑)	GPT	**0.118**	**0.242**	**0.500**	**0.396**	**0.320**
			Gemini	0.384	0.493	0.243	0.176	0.163
		DINO Image (↑)	GPT	0.140	0.288	0.254	0.223	0.176
			Gemini	0.241	0.396	0.380	0.404	0.320
		Avg	GPT	0.134	0.297	0.226	0.211	0.170
			Gemini	0.292	0.460	0.117	0.118	0.105
	Plausibility	CLIP Text (↑)	GPT	0.093	0.268	0.100	0.039	0.035
			Gemini	0.172	0.393	-0.137	-0.210	-0.154
		CLIP Image (↑)	GPT	0.092	0.228	0.109	0.184	0.147
			Gemini	0.216	0.346	0.184	0.387	0.320
		DINO Image (↑)	GPT	**0.073**	**0.222**	**0.483**	**0.447**	**0.345**
			Gemini	0.184	0.343	0.213	0.257	0.194
		Avg	GPT	0.086	0.239	0.231	0.223	0.176
			Gemini	0.191	0.361	0.087	0.145	0.120

Table 31: Results of Image Preservation using semantic similarity metrics (CLIP Text, CLIP Image, DINO Image) (ONLINE setting, refer to Figure 8 for prompt). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

	Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
IMAGE PRESERVATION	Unchanged Regions	CLIP Text	GPT	0.197	0.434	-0.076	-0.116	-0.096
			Gemini	0.168	0.397	-0.057	-0.103	-0.076
		CLIP Image	GPT	**0.166**	**0.258**	0.147	0.000	0.000
			Gemini	0.137	0.241	-0.037	-0.050	-0.040
		DINO Image	GPT	0.301	0.464	**-0.275**	**-0.318**	**-0.264**
			Gemini	0.208	0.365	0.044	0.006	0.011
		Avg	GPT	0.221	0.385	-0.068	-0.145	-0.120
			Gemini	0.171	0.334	-0.017	-0.049	-0.035
	Global Consistency	CLIP Text	GPT	0.197	0.434	-0.082	-0.116	-0.096
			Gemini	0.137	0.353	-0.087	-0.186	-0.150
		CLIP Image	GPT	**0.160**	**0.252**	**0.323**	**0.260**	**0.216**
			Gemini	0.083	0.194	0.102	0.118	0.102
		DINO Image	GPT	0.258	0.420	-0.062	0.010	0.008
			Gemini	0.186	0.346	-0.027	-0.002	0.020
		Avg	GPT	0.205	0.369	0.060	0.051	0.043
			Gemini	0.135	0.298	-0.004	-0.023	-0.009
	Identity Preservation	CLIP Text	GPT	**0.375**	**0.612**	**-0.146**	-0.082	-0.065
			Gemini	0.143	0.371	0.054	0.103	0.075
		CLIP Image	GPT	0.803	0.895	0.132	0.139	0.109
			Gemini	0.066	0.175	-0.160	-0.127	-0.109
		DINO Image	GPT	0.491	0.653	0.139	**0.167**	**0.131**
			Gemini	0.181	0.338	0.039	0.109	0.083
		Avg	GPT	0.556	0.720	0.042	0.075	0.058
			Gemini	0.130	0.295	-0.022	0.028	0.016

Table 32: Results of Edit Quality using semantic similarity metrics (CLIP Text, CLIP Image, DINO Image) (ONLINE setting, refer to Figure 8 for prompt). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

	Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
EDIT QUALITY	Scale Realism	CLIP Text	GPT	0.127	0.337	0.063	-0.042	-0.020
			Gemini	0.154	0.369	0.000	-0.034	-0.025
		CLIP Image	GPT	**0.069**	**0.175**	**0.169**	**0.333**	**0.254**
			Gemini	0.097	0.203	0.277	0.298	0.249
		DINO Image	GPT	0.157	0.311	0.124	0.276	0.209
			Gemini	0.184	0.338	0.104	0.214	0.162
		Avg	GPT	0.118	0.274	0.119	0.189	0.148
			Gemini	0.145	0.303	0.127	0.159	0.129
	Spatial Relationship	CLIP Text	GPT	0.125	0.318	-0.143	-0.174	-0.138
			Gemini	0.156	0.392	0.177	0.138	0.115
		CLIP Image	GPT	**0.040**	**0.137**	**0.486**	**0.585**	**0.471**
			Gemini	0.069	0.171	-0.121	-0.144	-0.115
		DINO Image	GPT	0.133	0.299	0.308	0.381	0.308
			Gemini	0.186	0.353	0.090	0.133	0.115
		Avg	GPT	0.099	0.251	0.217	0.264	0.214
			Gemini	0.137	0.305	0.049	0.042	0.038
	Texture and Detail	CLIP Text	GPT	0.123	0.302	-0.131	-0.091	-0.062
			Gemini	0.127	0.337	0.015	0.032	0.024
		CLIP Image	GPT	**0.123**	**0.259**	**0.404**	**0.215**	**0.164**
			Gemini	0.078	0.191	0.217	0.201	0.150
		DINO Image	GPT	0.188	0.372	-0.066	-0.036	-0.017
			Gemini	0.181	0.345	-0.051	-0.048	-0.036
		Avg	GPT	0.145	0.311	0.069	0.029	0.028
			Gemini	0.129	0.291	0.060	0.062	0.046
	Image Quality	CLIP Text	GPT	0.188	0.425	0.009	-0.016	-0.013
			Gemini	0.165	0.402	0.118	0.108	0.090
		CLIP Image	GPT	**0.135**	**0.227**	**0.308**	**0.222**	**0.184**
			Gemini	0.062	0.154	0.180	0.139	0.115
		DINO Image	GPT	0.250	0.412	-0.077	-0.098	-0.081
			Gemini	0.184	0.347	0.160	0.170	0.141
		Avg	GPT	0.191	0.355	0.080	0.036	0.030
			Gemini	0.137	0.301	0.153	0.139	0.115
	Color and Lighting	CLIP Text	GPT	0.117	0.296	-0.208	-0.213	-0.153
			Gemini	0.168	0.401	-0.222	-0.310	-0.230
		CLIP Image	GPT	**0.094**	**0.224**	**0.553**	**0.479**	**0.367**
			Gemini	0.134	0.247	0.480	0.461	0.361
		DINO Image	GPT	0.144	0.310	0.151	0.206	0.153
			Gemini	0.190	0.351	0.139	0.136	0.113
		Avg	GPT	0.118	0.277	0.165	0.157	0.122
			Gemini	0.164	0.333	0.132	0.096	0.081
	Seamlessness	CLIP Text	GPT	0.140	0.342	-0.047	0.019	0.021
			Gemini	0.161	0.382	-0.103	-0.212	-0.162
		CLIP Image	GPT	**0.123**	**0.249**	**0.544**	**0.421**	**0.322**
			Gemini	0.139	0.257	0.263	0.311	0.239
		DINO Image	GPT	0.165	0.345	0.157	0.213	0.163
			Gemini	0.194	0.362	0.098	0.114	0.091
		Avg	GPT	0.143	0.312	0.218	0.218	0.169
			Gemini	0.165	0.334	0.086	0.071	0.056

Table 33: Results of Instruction Fidelity using semantic similarity metrics (CLIP Text, CLIP Image, DINO Image) (ONLINE setting, refer to Figure 8 for prompt). Evaluation measures score prediction (MSE, MAE) and ranking quality (Pearson, Spearman, Kendall τ). Scores are reported for both GPT and Gemini.

	Factor	Metric	Model	MSE ↓	MAE ↓	Pearson ↑	Spearman ↑	Kendall τ ↑
INSTRUCTION FIDELITY	Alignment	CLIP Text	GPT	0.152	0.376	**0.141**	0.116	0.087
			Gemini	0.154	0.390	0.035	-0.053	-0.044
		CLIP Image	GPT	**0.075**	**0.173**	0.008	0.064	0.054
			Gemini	0.043	0.140	0.251	0.163	0.133
		DINO Image	GPT	0.190	0.346	0.042	**0.189**	**0.146**
			Gemini	0.174	0.342	0.182	-0.001	-0.006
		Avg	GPT	0.139	0.298	0.064	0.123	0.096
			Gemini	0.124	0.291	0.156	0.036	0.028
	Completeness	CLIP Text	GPT	0.152	0.376	**0.141**	0.116	0.087
			Gemini	0.157	0.394	0.140	0.139	0.116
		CLIP Image	GPT	**0.075**	**0.173**	0.008	0.064	0.054
			Gemini	0.034	0.126	0.334	0.290	0.240
		DINO Image	GPT	0.190	0.346	0.042	**0.189**	**0.146**
			Gemini	0.164	0.326	0.336	0.290	0.240
		Avg	GPT	0.139	0.298	0.064	0.123	0.096
			Gemini	0.118	0.282	0.270	0.240	0.199
	Plausibility	CLIP Text	GPT	0.137	0.344	0.161	0.086	0.062
			Gemini	0.192	0.422	-0.003	-0.103	-0.073
		CLIP Image	GPT	**0.096**	**0.204**	**0.211**	**0.274**	**0.219**
			Gemini	0.206	0.307	0.299	0.356	0.276
		DINO Image	GPT	0.177	0.329	0.065	0.119	0.095
			Gemini	0.212	0.375	0.170	0.233	0.181
		Avg	GPT	0.137	0.292	0.146	0.160	0.125
			Gemini	0.203	0.368	0.155	0.162	0.128