This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

I2I-Bench: A Comprehensive Benchmark Suite for Image-to-Image Editing Models

Juntong Wang*, Jiarui Wang*, Huiyu Duan*, Jiaxiang Kang, Guangtao Zhai, Xiongkuo Min†
Institute of Image Communication and Network Engineering
Shanghai Jiao Tong University, Shanghai, China

Abstract

Image editing models are advancing rapidly, yet comprehensive evaluation remains a significant challenge. Existing image editing benchmarks generally suffer from limited task scopes, insufficient evaluation dimensions, and heavy reliance on manual annotations, which significantly constrain their scalability and practical applicability. To address this, we propose ***I2I-Bench****, a comprehensive benchmark for image-to-image editing models, which features* ***(i) diverse tasks****, encompassing 10 task categories across both single-image and multi-image editing tasks,* ***(ii) comprehensive evaluation dimensions****, including 30 decoupled and fine-grained evaluation dimensions with automated hybrid evaluation methods that integrate specialized tools and large multimodal models (LMMs), and* ***(iii) rigorous alignment validation****, justifying the consistency between our benchmark evaluations and human preferences. Using I2I-Bench, we benchmark numerous mainstream image editing models, investigating the gaps and trade-offs between editing models across various dimensions. The dataset and code are released at https://github.com/IntMeGroup/I2I-Bench.*

1. Introduction

Image editing has long been important tasks in computer vision [9, 17, 40, 41]. Driven by the advancements of large multimodal models (LMMs) [1, 7, 8, 18, 25, 35, 50], image editing methods have made rapid progress in the past few years, shifting from traditional local operations to instruction-induced content modification [3, 54]. Moreover, the capabilities of editing models have also rapidly expanded, achieving not only single-image editing (SE) [3, 5, 6, 28, 54] but also more challenging multiple-image-editing (ME) [9, 46]. The rapid evolution across tasks and modalities underscores the urgent need for a new comprehensive benchmarks designed to rigorously assess these advanced capabilities.

However, existing evaluation paradigms exhibit clear shortcomings, struggling to keep pace with the rapid advancement of editing model capabilities. First, traditional metrics [11, 12, 37] like PSNR or LPIPS [52] are fundamentally inadequate, as they primarily assess pixel-level similarity and completely fail to evaluate whether complex semantic edits have been correctly executed. Though some more recent evaluation benchmarks [2, 15, 23, 33, 34, 48] have attempted to address this, they remain limited in both scope and granularity. Some benchmarks [15] possess a very limited test scope [11, 16, 32, 42, 43], while others [10, 19, 45, 47, 53] lack the independent, quantitative analysis in terms of fine-grained dimensions which are crucial for editing model understanding and optimization. Furthermore, in terms of evaluation methods, current benchmarks face significant limitations. Some rely on manual annotation methods that are costly, non-scalable, and difficult-to-reproduce. Other approaches, such as score-based training methods [48], still heavily rely on these manual annotations. Meanwhile, many automated methods perform poorly when handling complex semantic and cognitive tasks. More importantly, despite the emergence of multi-image editing task [9, 40, 41, 46], existing evaluation frameworks almost entirely lack coverage of such challenging tasks. A comprehensive and effective image editing evaluation suite is still lacking.

To systematically address all the aforementioned limitations, we propose **I2I-Bench**, a comprehensive benchmark suite for evaluating image editing models. An overview of the entire benchmark suite is shown in Figure 1. I2I-Bench aims to provide a comprehensive, automated, fine-grained, and human preference-aligned evaluation framework. Firstly, our benchmark contains **diverse tasks** to facilitate comprehensive evaluation. Specifically, a total of 1000 prompts, systematically covering 10 carefully designed task categories across both from single-image and multi-image editing tsaks. Secondly, I2I-Bench contains comprehensive evaluation dimensions. We propose a hierarchical, decoupled evaluation framework with 30 fine-grained dimensions. For each dimension, we define a reproducible, automated hybrid evaluation pipeline, integrating specialized "specialist" tools for precision score predic-

*Equal contribution
†Corresponding author

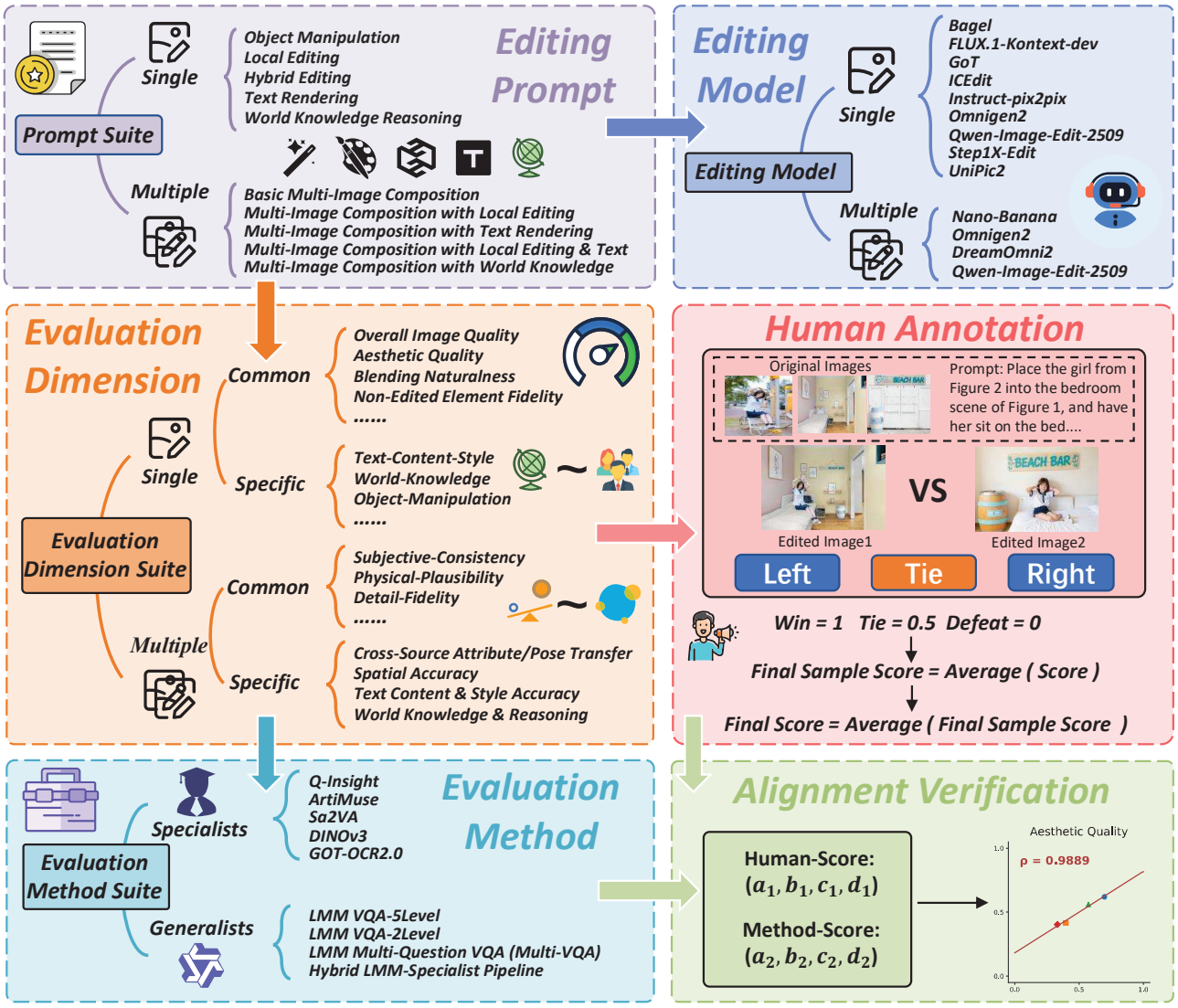


Figure 1. An overview of the proposed image-to-image editing evaluation benchmark suite, I2I-Bench. The process starts with our large-scale **Prompt Suite**, which defines the editing tasks. These prompts are fed into the **Editing Model** to edit images. The prompts also guide the selection of relevant dimensions from our hierarchical **Evaluation Dimension** Suite. Each dimension, in turn, specifies both the automated **Evaluation Method** Suite (combining Specialists and Generalists) and the criteria for **Human Annotation**. Finally, the results from the automated methods and human annotations are compared for **Alignment Verification** to ensure the reliability of our benchmark.

tion and general "generalists" for semantic understanding. Thirdly, we conduct rigorous alignment validation for I2I-Bench. We conduct large-scale human preference experiments to validate our automated metrics. This validation confirms that our automated evaluation results, particularly those from our hybrid pipeline and LMM-based evaluators, achieve an extremely high consistency with human judgment, ensuring the reliability of our benchmark. Finally, I2I-Bench provides **in-depth benchmarking and actionable insights**. Using this comprehensive suite, we evaluate numerous mainstream image editing models. This analysis reveals key performance trade-offs, identifies critical bottlenecks (such as failures in physical plausibility and abstract reasoning), and exposes universal limitations in current models. These findings provide actionable insights to guide future research toward addressing these fundamental gaps.

2. Related Work

2.1. Image Editing Models

Image editing models have undergone rapid development in recent years. Early models such as Instruct-Pix2Pix [3] introduce instruction-based editing. Subsequent works such as Emu-Edit [28] and Qwen-Image-Edit [40] focus on improving editing fidelity and blending quality. Recently, the emergence of multi-reference editing (ME) models, represented by Nano-Banana [9] and DreamOmni2 [46], have expanded

the scope of possible edits to include more complex tasks such as cross-source attribute transfer and subject consistency. This rapid iteration of model capabilities underscores the growing need for more sophisticated evaluation methodologies [13].

2.2. Image Editing Evaluation Benchmarks

Existing evaluation benchmarks [14, 15, 24, 26, 27, 34, 48] exhibit several limitations in their evaluation coverage of current model capabilities. For instance, TedBench [15] has a very limited test scope. EditBench [34] covers multiple editing types but relies heavily on expensive and hard-to-reproduce manual annotations. To achieve automation, EditVal [2] adoptes LMMs [51] for evaluation but struggles with processing complex semantic and cognitive tasks. LMM4Edit [48] is also based on LMMs [1], but its evaluation dimensions remain restricted to a limited set of aspects such as perceptual quality, editing alignment, and attribute preservation. In summary, there remains a clear gap due to the lack of a comprehensive benchmark capable of fine-grained, automated evaluation of advanced cognitive tasks, especially for emerging multi-image editing tasks.

3. The I2I-Bench Suite

I2I-Bench is a comprehensive benchmark suite composed of three integral components for the systematic evaluation of image editing models [40]: 1) a large-scale and structured Prompt Suite that provides diverse test cases; 2) a hierarchical Evaluation Dimension & Method Suite that defines fine-grained aspects of editing quality with the corresponding quantitative methods; and 3) a Human Preference Annotation protocol designed to validate the alignment of our automated metrics with human judgment. An overview of the entire benchmark structure is shown in Figure 1.

3.1. Prompt Suite

The I2I-Bench Prompt Suite contains 1000 meticulously crafted prompts, equally split between single-image and multi-image editing tasks. This structured collection provides broad and systematic coverage of common editing scenarios and challenges. Illustrative examples for each prompt category are shown in Figure 2.

Single-Image Editing Prompts. This subset includes 500 prompts, structured as 100 base scenarios, each providing 5 prompts corresponding to the 5 SE categories, testing distinct aspects of single-image editing: *Object Manipulation* (assessing fundamental global edits like adding, removing, or replacing objects); *Local Editing* (evaluating precision in local attribute modification); *Hybrid Editing* (challenging models with complex, multi-part instructions requiring both global and local changes); *Text Rendering* (focusing on text accuracy); and *World Knowledge & Reasoning* (probing the application of external knowledge).

Multi-Image Editing Prompts. This subset contains 500 prompts, with 100 base scenarios covering 5 task categories of increasing complexity: *Basic Combination* (testing simple subject extraction and composition); *Combination + Content Editing* (requiring subsequent attribute/pose edits after combination); *Combination + Text Rendering* (adding text to a composite scene); *Combination + Editing + Text* (involving multi-step, mixed instructions); and *Combination + Reasoning* (demanding integration of external knowledge based on the combined context).

Relationship Between Prompts and Dimensions. Our prompt categories (organized by user task scenarios) and evaluation dimensions (a fine-grained decomposition of quality) have a deliberate "many-to-many" design. This allows a single, complex prompt (*e.g.,* from "Hybrid Editing") to holistically assess multiple fine-grained quality aspects (*e.g.,* "Blending Naturalness", "Object Manipulation Accuracy", and "Physical Plausibility"), simultaneously within an integrated task, facilitating integrated quality assessment beyond isolated dimension testing.

3.2. Evaluation Dimension and Method Suite

To comprehensively evaluate the diverse and complex editing tasks defined in the Prompt Suite, we design a hierarchical Evaluation Dimension and Method Suite. Traditional image quality metrics (such as PSNR or LPIPS [52]) only measure superficial pixel similarity and fail to assess the semantic accuracy, physical plausibility, or logical coherence in editing outputs. In contrast, our framework moves beyond low-level fidelity to evaluate deep semantic alignment in image editing. Our evaluation suite comprises a total of 30 fine-grained evaluation dimensions, divided equally into 15 Single-Reference dimensions and 15 Multi-Reference dimensions. Within each category, the dimensions are further divided into "Common Dimensions", applicable across all tasks for assessing fundamental quality, and "Specific Dimensions", tailored to specific instruction types or editing scenarios. Crucially, we define a reproducible evaluation pipeline for each dimension, integrating two categories of tools: "Specialists" and "Generalists". Specialists refer to mature, quantitative tools trained for specific tasks (*e.g.,* Q-Insight [20], ArtiMuse [4], DINOv3 [29]). Generalists refer to powerful LMMs (Qwen3-VL-8B-Instruct [1]), which we leverage to evaluate dimensions requiring complex semantic understanding, world knowledge, and reasoning. In the following, we present the definitions of the dimensions integrated with their corresponding evaluation methods.

3.2.1. Single-Reference Editing Dimensions and Methods

The 15 single-reference dimensions are categorized into 7 Common Dimensions and 8 Specific Dimensions.

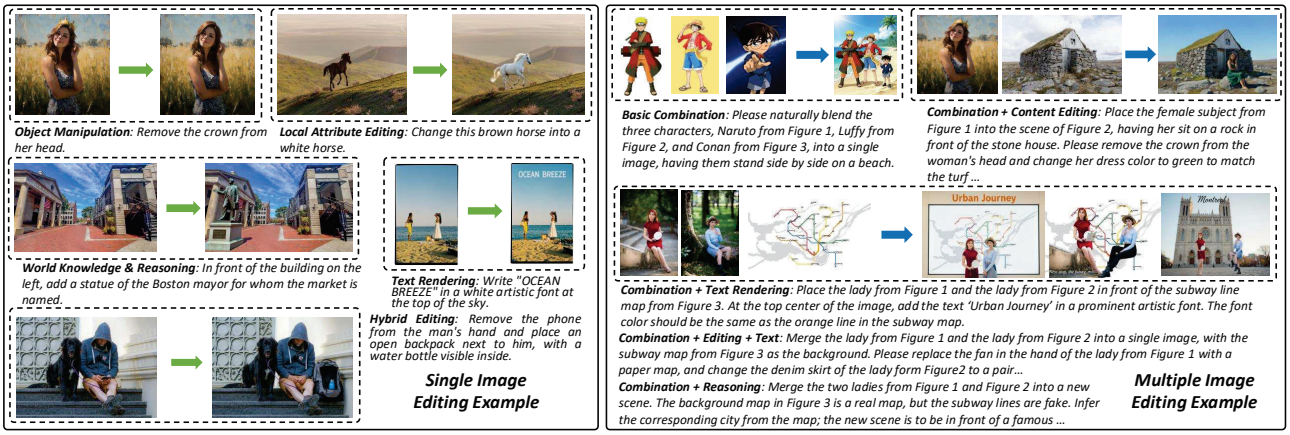


Figure 2. Visualization of the 10 task categories in the I2I-Bench Prompt Suite. The left half shows 5 single-image editing (SE) tasks, from "Object Manipulation" to "World Knowledge & Reasoning". The right half shows 5 multi-image editing (ME) tasks, illustrating increasing complexity from "Basic Combination" to "Combination + Reasoning".

A. Common Dimensions

Common dimensions assess the fundamental quality of all single-reference editing tasks.

Overall Image Quality and **Aesthetic Quality.** We evaluate the objective technical quality (*e.g.,* clarity, noise) and the artistic appeal of the image, respectively. We use specialist models Q-Insight [20] and ArtiMuse [4] to obtain direct quantitative scores for these two dimensions, respectively.

Blending Naturalness and **Generative/Editing Artifacts.** We evaluate the smoothness of the transition between the edited region and its surroundings, and the introduction of Artificial Intelligence (AI)-specific flaws (*e.g.,* broken structures, unnatural textures). We employ the LMM VQA-5Level pipeline [31, 44], a method based on Visual Question Answering (VQA). The LMM [1] is asked to choose from a 5-level scale {"excellent", "good", "fair", "poor", "bad"}, where each choice c_i is assigned a weight $w_i \in \{1, 0.75, 0.5, 0.25, 0\}$. The LMM's logits z_i are converted to probabilities P_i via the Softmax function. The final score is the weighted average:

$$\text{Score}_{\text{5-level}} = \sum_{i=1}^{5} w_i \cdot P_{\text{LMM}}(c_i | I, Q), \tag{1}$$

where I represents the required image(s) for evaluation (*e.g.,* I_{gen}, or the $I_{\text{orig}}, I_{\text{gen}}$ pair), and Q is the question or instruction. For Editing Artifacts, the LMM receives I_{orig}, I_{gen}, and Q, while for Blending Naturalness, it focuses on the integration in I_{gen}.

Instruction Following (Macro). We assess the model's overall understanding and execution of the instruction's core intent [30]. We reuse the LMM VQA-5level pipeline (Eq. (1)) to evaluate the high-level semantic grasp.

Non-Edited Region Fidelity. We assess whether image regions not targeted by the instruction remain unchanged post-edit. We use a hybrid LMM-specialist pipeline. This process combines LMM semantics with specialist precision: (1) The LMM receives the original/generated images and instruction, generating a segmentation command (*e.g.,* "segment the horse"). (2) The specialist segmentation model Sa2VA [49] uses this command to generate corresponding masks for the edited region, M_{edit}. (3) We invert this mask ($M_{\text{non-edit}} = \text{NOT}(M_{\text{edit}})$) to obtain the non-edited region mask. (4) The specialist feature extractor DINOv3 [29] extracts feature vectors f_{orig} and f_{gen} from within $M_{\text{non-edit}}$. (5) The final score is the cosine similarity between these vectors:

$$\text{Score}_{\text{Hybrid-Sim}} = \frac{f_{\text{orig}} \cdot f_{\text{gen}}}{||f_{\text{orig}}|| \cdot ||f_{\text{gen}}||}. \tag{2}$$

Physical Plausibility. We evaluate whether the edit adheres to physical laws [36] (*e.g.,* lighting, perspective, gravity). We employ the LMM multi-question VQA (Multi-VQA) pipeline. This is a two-stage process: (1) Question Generation: The LMM receives the original image(s) and instruction, and is prompted to design N closed-ended (Yes/No) questions for the specific dimension (*e.g.,* "Is the shadow direction of the new object correct?"). (2) Question Answering: The LMM receives the generated questions Q_i, original image(s), edited image, and instruction, and answers 'Yes' or 'No' to each. The final score is the ratio of 'Yes' answers:

$$\text{Score}_{\text{Multi-VQA}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\text{LMM}(I_{\text{gen}}, I_{\text{orig}}, Q_i) = \text{'Yes'}). \tag{3}$$

B. Specific Dimensions

These 8 dimensions evaluate the execution quality of specific instructions.

Object Manipulation Accuracy, Local Attribute Accuracy, Action/State Change Accuracy, Spatial Accuracy.

We evaluate task-specific accuracy, such as the correctness of "add/remove/replace" object operations, the accurate application of "change color/material" attribute modifications, and the faithful execution of spatial descriptions (*e.g.,* "to the left of"). We primarily use the LMM VQA-2level pipeline [21]. The LMM is asked a binary (Yes/No) question regarding the success of the specific editing operation (*e.g.,* "Has the horse's color been successfully changed to white?"). The final score is the probability of the 'Yes' answer:

$$\text{Score}_{\text{2-level}} = P_{\text{LMM}}(\text{'Yes'}|I, Q). \tag{4}$$

Text Content & Style Accuracy. We assess the model's ability to render text in images, with correct spelling and appropriate visual presentation. We use a multi-step hybrid pipeline: (1) Content Accuracy: We use the specialist OCR model GOT-OCR2.0 [38] to compute S_c based on its distance (*e.g.,* Levenshtein) from the target text. (2) Style and Position: We use the LMM VQA-5level pipeline (Eq. (1)), instructing the LMM to ignore spelling and provide a 1-5 discrete score (S_s) based on style and position match. (3) Score Fusion: The final score ($\text{Score}_{\text{final}}$) is calculated by combining S_s (style score) and S_c (content accuracy) as follows:

$$\text{Score}_{\text{final}} = \left(\frac{S_s - 1}{4}\right) \times \begin{cases} 1.0 & \text{if } S_c = 1.0 \\ 0.8 & \text{if } 0.8 \leq S_c < 1.0 \\ 0.5 & \text{if } 0.6 \leq S_c < 0.8 \\ 0.1 & \text{if } S_c < 0.6. \end{cases} \tag{5}$$

World Knowledge & Reasoning. We evaluate instructions requiring external knowledge (*e.g.,* "add a statue of ..."). We use the LMM VQA-2level pipeline [21] (Eq. (4)) to judge if the generated image matches a pre-defined correct answer set.

Subject Identity Fidelity. We assesses whether a subject's core identity is preserved during attribute modifications. We still use the hybrid LMM-specialist pipeline (Eq. (2)) to compute feature similarity between the original and edited images, but we only compare the segmented subject regions directly (without mask inversion).

Composition & Interaction. We evaluate the logical interaction and visual coherence between newly introduced or altered elements and existing scene content. We reuse the LMM multi-question VQA pipeline (Eq. (3)) to ask specific questions about interaction plausibility.

3.2.2. Multi-Reference Editing Dimensions and Methods

The 15 multi-reference dimensions consist of 9 shared dimensions (reused from single-reference, with identical evaluation methods) and 6 multi-reference-specific dimensions. These specific dimensions focus on evaluating cross-image composition and consistency.

Non-Edited Element Fidelity. Assesses whether elements from source images that are not designated as editing targets (*e.g.,* background elements) are erroneously modified or discarded during composition. We reuse the Hybrid LMM-Specialist pipeline (Eq. 2). The LMM segments elements to be preserved, inverts the corresponding mask, then DINOv3 [29] is applied to compute similarity on the non-edited elements.

Subject Consistency and **Detail Fidelity/Preservation.** We assess the preservation of a subject's identity and fine details from its source image to the newly generated image. We reuse the hybrid LMM-specialist pipeline (Eq. (2)) to compute feature similarity, by directly comparing the segmented subjects from the source and edited images (without mask inversion).

Subject Extraction & Composition. We evaluate whether subjects are completely extracted and correctly composed into the new scene. We implement a two-step multiplicative scoring process: (1) The LMM VQA-2level pipeline [21] (Eq. (4)) provides a binary score ($\text{Score}_{\text{count}}$) for the correct number of subjects. (2) Then multiplied by the $\text{Score}_{\text{consistency}}$ (calculated via the **Subject Consistency** hybrid pipeline) to yield $\text{Score}_{\text{final}} = \text{Score}_{\text{count}} \times \text{Score}_{\text{consistency}}$.

Inter-Subject Consistency and **Cross-Source Attribute/Pose Transfer.** We assess the visual coherence of subjects from different sources when composed together (e.g., in lighting, scale, and style); or evaluate the fidelity of transferring attributes or poses between subjects across sources. We still use the LMM VQA-2level (Eq. (4)) or Multi-VQA (Eq. (3)) pipelines to pose specific questions about these complex cross-image interactions.

3.3. Human Preference Annotation

To validate the alignment of our proposed automated evaluation dimensions (especially those relying on LMMs) with genuine human perception, we conduct a large-scale human preference annotation experiment, following established practices in benchmark development.

Model and Sample Selection. We evaluate 9 single image editing and 4 multi-image editing models. For ME tasks, all 4 models are compared in a pairwise manner. For SE tasks, to manage annotation costs, we randomly assign a unique combination of 4 models for comparison within each of the 15 SE dimensions. We uniformly sample 85 items (510 pairs) for all 30 evaluation dimensions to ensure broad and representative coverage.

Annotation Process. We use a pairwise comparison (A vs. B vs. Tie) format. For each sample, annotators are provided with the source image(s), instruction, and a clear definition of the specific dimension being evaluated. They are strictly

Table 1. Results of the single-image editing benchmark. All scores are normalized; higher is better. Best scores are highlighted.

Model	Action-State-Change	Aesthetic-Quality	Blending-Naturalness	Composition-Interaction	Editing-Artifacts	Image-Quality	Instruction-Following	Local-Attribute
Qwen-Image-Edit-2509 [40]	**0.977**	0.589	**0.807**	**0.823**	**0.907**	0.811	**0.930**	**0.961**
Step1X-Edit [22]	0.819	0.584	0.738	0.804	0.838	0.807	0.850	0.895
UniPic-2 [39]	0.964	0.567	0.733	0.819	0.798	0.799	0.862	0.941
Bagel [5]	0.781	0.584	0.669	0.752	0.756	0.810	0.785	0.865
FLUX.1-Kontext-dev [17]	0.882	**0.592**	0.697	0.753	0.753	**0.823**	0.787	0.851
Omnigen2 [41]	0.844	0.569	0.602	0.750	0.694	0.812	0.755	0.848
ICEdit [54]	0.722	0.590	0.599	0.734	0.623	0.797	0.680	0.784
GoT [6]	0.659	0.563	0.431	0.621	0.444	0.786	0.490	0.606
instruct-pix2pix [3]	0.343	0.552	0.176	0.542	0.173	0.769	0.221	0.404

Model	Non-Edited-Fidelity	Object-Manipulation	Physical-Plausibility	Spatial-Accuracy	Subject-Identity	Text-Content-Style	World-Knowledge	Overall
Qwen-Image-Edit-2509 [40]	0.898	**0.937**	0.527	**0.951**	0.768	**0.666**	**0.749**	**0.813**
Step1X-Edit [22]	0.928	0.862	0.537	0.844	0.799	0.456	0.655	0.773
UniPic-2 [39]	0.818	0.927	**0.562**	0.930	0.677	0.372	0.605	0.767
Bagel [5]	**0.938**	0.838	0.502	0.808	**0.830**	0.438	0.668	0.742
FLUX.1-Kontext-dev [17]	0.852	0.804	0.479	0.813	0.694	0.404	0.570	0.727
Omnigen2 [41]	0.891	0.810	0.488	0.793	0.764	0.434	0.532	0.713
ICEdit [54]	0.902	0.715	0.518	0.732	0.781	0.265	0.446	0.684
GoT [6]	0.897	0.596	0.399	0.500	0.760	0.038	0.469	0.575
instruct-pix2pix [3]	0.742	0.285	0.348	0.247	0.647	0.014	0.300	0.416

Table 2. Results of the multi-image editing benchmark. All scores are normalized; higher is better. Best scores are highlighted.

Model	Aesthetic-Quality	Blending-Naturalness	Composition-Interaction	Cross-Source-Attribute	Detail-Fidelity	Image-Quality	Instruction-Following	Inter-Subject-Consistency
Nano-Banana [9]	**0.580**	**0.562**	**0.693**	**0.700**	**0.514**	0.803	**0.830**	**0.582**
Qwen-Image-Edit-2509 [40]	0.566	0.556	0.665	0.628	0.506	0.800	0.744	0.564
DreamOmni2 [46]	0.555	0.579	0.604	0.510	0.478	**0.827**	0.658	0.556
Omnigen2 [41]	0.531	0.546	0.587	0.518	0.462	**0.827**	0.620	0.543

Model	Non-Edited-Fidelity	Physical-Plausibility	Spatial-Accuracy	Subject-Consistency	Subject-Extraction	Text-Content-Style	World-Knowledge	Overall
Nano-Banana [9]	**0.655**	**0.433**	**0.894**	**0.535**	**0.530**	**0.687**	**0.721**	**0.636**
Qwen-Image-Edit-2509 [40]	0.613	0.424	0.893	0.516	0.505	0.660	0.044	0.604
DreamOmni2 [46]	0.609	0.388	0.762	0.472	0.434	0.323	0.210	0.562
Omnigen2 [41]	0.564	0.346	0.749	0.463	0.445	0.479	0.038	0.545

instructed to judge only on that single dimension, disregarding all other aspects. For example, when evaluating Subject Identity Fidelity, annotators are instructed to select the image that better preserves subject identity, even if it exhibits noticeable artifacts in Blending Naturalness artifacts.

Win Ratio Calculation. We calculate a Win Ratio for each model per dimension based on pairwise results. In this calculation, a 'win' is assigned a numerical value of 1, a 'tie' is assigned 0.5, and a 'loss' is assigned 0. The final Win Ratio for each model in a specific dimension is calculated as the total score accumulated divided by the number of comparisons it participated for that dimension.

4. Experiments

In this section, we conduct a comprehensive evaluation of a series of advanced image editing models using the proposed I2I-Bench. We first present the detailed performance of all models across 30 evaluation dimensions in Subsection 4.1. We then validate the consistency between the I2I-Bench evaluation methodology and human perception through large-scale human preference annotation. Finally, we provide a deeper analysis by task category in Subsection 4.4 and a comparative analysis between single-image and multi-image editing models in Subsection 4.5.

4.1. Evaluation by Dimension

We first present the main evaluation results of I2I-Bench across all 30 fine-grained dimensions. The detailed scores for all nine SE models and four ME models are presented in Table 1 and Table 2, respectively. Key performance trends and trade-offs are visualized in Figure 3.

For SE models, as illustrated in Figure 3(a)-(b), Qwen-Image-Edit-2509 [40] excels in "Blending Naturalness" and "Editing Artifacts", and leads in most task execution dimensions (*e.g.,* "Instruction-Following-Macro"). We also identify clear model-specific trade-offs: Bagel [5], for instance, attains high score on "Non-Edited-Element-Fidelity" but low on "Blending Naturalness", highlighting a conflict between preserving the background and blending new content.

For ME models, Figure 3(c-d) reveals that these tasks pose substantially greater challenges, with foundational quality scores (*e.g.,* "Aesthetic Quality") being generally lower than those in SE tasks. Nano-banana [9] demonstrates strong

Table 3. **Human Preference Alignment and Ablation Study.** This table presents the Pearson correlation (ρ) coefficients between I2I-Bench automated metrics and human preferences (Win Ratio). The results show extremely high consistency across all 30 dimensions, strongly validating the reliability of our evaluation methodology. The table also presents an ablation study comparing our hybrid **I2I-Bench** pipeline vs. a **Pure LMM-Baseline** (general LMM VQA-5level), which validates our hybrid design.

Metric (SE)	Action-State-Change	Aesthetic-Quality	Blending-Naturalness	Composition-Interaction	Editing-Artifacts	Image-Quality	Instruction-Following	Local-Attribute
Ours (ρ)	**0.9839**	**0.9889**	**0.8866**	**0.9997**	**0.9006**	**0.9033**	**0.9960**	**0.9877**
LMM-Baseline (ρ)	0.9769	-0.0643	0.8866	0.9851	0.8983	0.4154	0.9960	0.9842
Metric	Non-Edited-Fidelity	Object-Manipulation	Physical-Plausibility	Spatial-Accuracy	Subject-Identity	Text-Content-Style	World-Knowledge	**Overall**
Ours (ρ)	**0.9019**	**0.9787**	**0.8055**	**0.9303**	**0.9133**	**0.9979**	**0.9628**	**0.9425**
LMM-Baseline (ρ)	0.7687	0.9673	0.6958	0.9191	-0.3494	0.8847	0.9506	0.7277
Metric (ME)	Aesthetic-Quality	Blending-Naturalness	Composition-Interaction	Cross-Source-Attribute	Detail-Fidelity	Image-Quality	Instruction-Following	Inter-Subject-Consistency
Ours (ρ)	**0.9034**	**0.6177**	**0.8951**	**0.9917**	**0.8767**	**0.6579**	**0.9248**	**0.9469**
LMM-Baseline (ρ)	0.7138	0.6177	0.8205	0.9902	0.8092	-0.9202	0.9248	0.8293
Metric	Non-Edited-Fidelity	Physical-Plausibility	Spatial-Accuracy	Subject-Consistency	Subject-Extraction	Text-Content-Style	World-Knowledge	**Overall**
Ours (ρ)	**0.6459**	**0.9017**	**0.9008**	**0.9401**	**0.9732**	**0.9043**	**0.9523**	**0.8688**
LMM-Baseline (ρ)	0.5978	0.8887	0.8230	0.6616	0.7763	0.8573	0.8534	0.6829

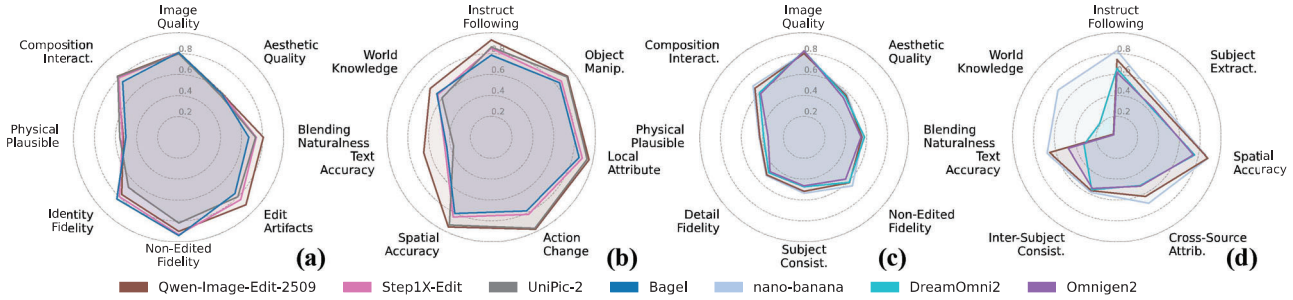


Figure 3. Capability radar charts for the evaluated models on key dimensions. (a) Foundational Quality & Fidelity (SE models). (b) Task Execution & Advanced Capabilities (SE models). (c) Foundational Quality & Fidelity (ME models). (d) Task Execution & Advanced Capabilities (ME models).

overall performance, particularly in "World-Knowledge-Reasoning", where other models like Omnigen2 [41] and Qwen-Image-Edit-2509 [40] completely fail. This significant disparity, which highlights a fundamental gap in reasoning capabilities, is analyzed in detail in Section 4.5.

4.2. Validation of Human Alignment

To verify the reliability of I2I-Bench and its alignment with human perception, we conduct a large-scale human annotation experiment, as described in Subsection 4.2. We adopt a pairwise comparison format and calculate a human preference "Win Ratio" for each model in each dimension. We then compute the Pearson correlation coefficient (ρ) between the I2I-Bench automated evaluation scores and the human preference win ratios. As shown in Table 3 (see the "Ours (ρ)" rows). we observe strong consistency across all 30 dimensions, validating the effectiveness of I2I-Bench.

4.3. Ablation Study on Evaluation Pipeline

To validate the necessity of our hybrid evaluation paradigm, we conduct an ablation study comparing our method against a "Pure LMM" baseline that uses the LMM VQA-5level as the general-purpose evaluation method. As presented in Table 3, a direct comparison between the "Ours (ρ)" and "LMM-Baseline (ρ)" rows strongly supports the design of I2I-Bench: the LMM baseline fails catastrophically on dimensions requiring specialist perception (*e.g.*, Aesthetic-Quality $\rho = -0.0643$ vs. our 0.9889), proving our "specialist" tools are indispensable. Simultaneously, our designed "hybrid LMM-specialist" pipeline also significantly outperforms the LMM baseline on complex compositional tasks. For specific accuracy dimensions, our selected LMM VQA-2level method offers superior interpretability—its probability output ($P(\text{'Yes'})$) allows for an intuitive 0.5 threshold to classify an operation as "successful," a property the 5-level score lacks. This ablation study confirms that I2I-Bench's hybrid design is critical for achieving robust, interpretable, and highly human-aligned evaluation.

4.4. Evaluation by Task Category

The I2I-Bench Prompt Suite is systematically organized into distinct task categories. To identify model-specific limi-

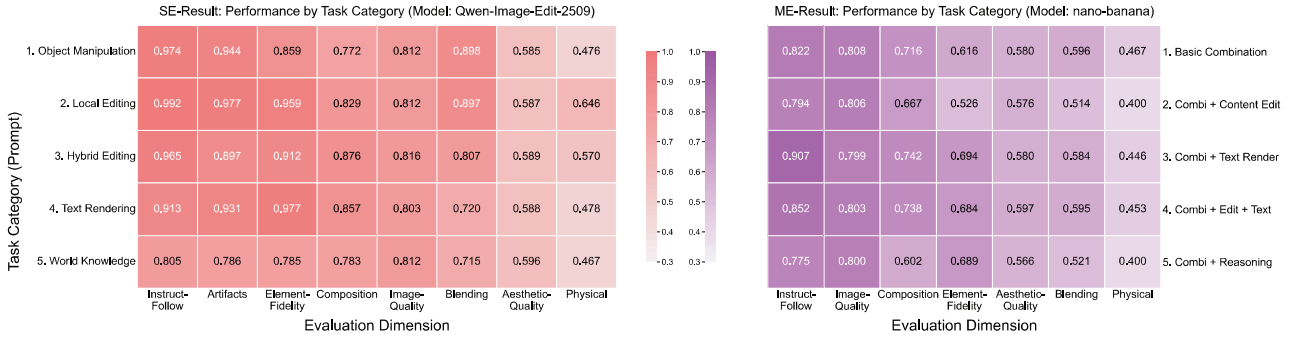


Figure 4. Performances of top-performing SE and ME models on common dimensions across task categories. (1) The performance of Qwen-Image-Edit-2509 (SE) as task cognitive complexity increases. (2) The performance of nano-banana (ME) varies across complex combination tasks.

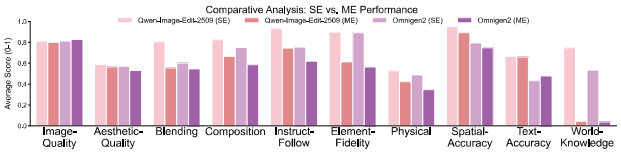


Figure 5. Performance comparison between Single-Image Editing (SE) and Multi-Image Editng (ME) tasks for Qwen-Image-Edit-2509 and Omnigen2 on shared dimensions.

tations, we analyze the average performance on common dimensions across these categories. As illustrated in Figure 4, for the top-performing SE model Qwen-Image-Edit-2509 [40], foundational quality ("Image-Quality") remains stable, but the cognitive complexity of the task significantly impacts key capabilities. This trend is most evident in the "Instruction-Following-Macro" dimension, showing a steady decline from "Local Editing" (0.992) to "World Knowledge" (0.805). A similar decrease is observed in "Blending-Naturalness" (from 0.897 to 0.715), indicating that as the task's cognitive load increases, the model's ability to follow instructions and maintain fidelity degrades.

For the top-performing ME model nano-banana, as illustrated in Figure 4, we again observe stable "Image-Quality" but identify two distinct bottlenecks. First, a "semantic bottleneck" emerges in the "Combi + Content Edit" task, the model achieves its lowest scores in fidelity-related dimensions such as "Blending-Naturalness" (0.514) and "Physical-Plausibility" (0.400). This suggests the complex semantic task of "combining and then deeply modifying" is a key challenge. Second, a "logic bottleneck" is evident in the "Combi + Reasoning" task, the model performs worst in "Composition-Interaction" (0.602) and "Instruction-Following-Macro" (0.775), indicating that abstract logic requirements hinder instruction understanding.

4.5. Comparative Analysis: SE vs. ME

Qwen-Image-Edit-2509 [40] and Omnigen2 [41] support both SE and ME tasks, allowing for a direct comparison of the challenges posed by multi-image inputs. As shown in Figure 5, both models exhibit a general performance degradation when shifting from SE to ME tasks. For Qwen-Image-Edit-2509, its "Blending-Naturalness" score drops from 0.807 (SE) to 0.556 (ME), and "Instruction-Following-Macro" drops from 0.930 (SE) to 0.744 (ME). These results indicate that processing and fusing information from multiple image sources impose significantly higher demands on model capabilities.

However, we must distinguish between performance drops on comparable metrics and fundamental increases in task difficulty. First, some metrics have different nuances (*e.g.,* "Non-Edited Fidelity" in SE refers to "Region" fidelity, whereas in ME it refers to "Element" fidelity). Second, the most dramatic difference is in the "World-Knowledge-Reasoning" dimension, where both models' scores plummet from acceptable SE levels (Qwen-Image-Edit-2509: 0.749, Omnigen2: 0.533) to near-zero (Qwen-Image-Edit-2509: 0.044, Omnigen2: 0.038). This is not a "capability loss" but reflects our intentional benchmark design. To test the upper limits of advanced models, our ME prompts for this dimension involve high-difficulty abstract reasoning (*e.g.,* Sudoku, map recognition) far more complex than the knowledge retrieval tasks in SE. This result demonstrates that the "knowledge retrieval" capability shown in SE tasks does not extend to the "abstract logical reasoning" required in our ME tasks, revealing a significant capability gap.

5. Conclusion

We propose I2I-Bench, a comprehensive image-to-image editing benchmark [9, 40], featuring a systematic prompt suite (10 categories) and a hybrid "Specialist-Generalist" evaluation system (30 dimensions). Our large-scale study validates its high consistency with human judgment and the necessity of this hybrid design. Using this benchmark, we reveal key cognitive trade-offs and universal limitations in current models, particularly failures in Physical Plausibility, multi-lingual Text Rendering, and Abstract Logical Reasoning. I2I-Bench provides a valuable tool to guide future research in addressing these gaps.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62522116, 62271312, 62132006, 62401365, 62225112, and U24A20220, in part by STCSM under Grant 22DZ2229005, and in part by the China Postdoctoral Science Foundation under Grants BX20250411 and 2025M773473.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1, 3, 4

[2] Samyadeep Basu, Mehrdad Saberi, Shweta Bhardwaj, Atoosa Malemir Chegini, Daniela Massiceti, Maziar Sanjabi, Shell Xu Hu, and Soheil Feizi. Editval: Benchmarking diffusion based text-guided image editing methods. *arXiv preprint arXiv:2310.02426*, 2023. 1, 3

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2023. 1, 2, 6

[4] Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, Bo Qu, Wenhai Wang, Yu Qiao, Dajuin Yao, and Yihao Liu. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. *arXiv preprint arXiv:2507.14533*, 2025. 3, 4

[5] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 1, 6

[6] Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, Xihui Liu, and Hongsheng Li. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 1, 6

[7] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, et al. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025. 1

[8] Google. https://ai.google.dev/gemini-api, 2025. Accessed: 2025-10-13. 1

[9] Google. Gemini-2.5-flash-image. https://aistudio.google.com/, 2025. Accessed: 2025-10-13. 1, 2, 6, 8

[10] Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and Chongyi Li. EvalMuse-40K: A Reliable and Fine-Grained Benchmark with Comprehensive Human Annotations for Text-to-Image Generation Model Evaluation. *arXiv preprint arXiv:2412.18150*, 2024. 1

[11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7514–7528, 2021. 1

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, page 6629–6640, 2017. 1

[13] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, et al. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[14] Bohan Jia, Wenxuan Huang, Yuntian Tang, Junbo Qiao, Jincheng Liao, Shaosheng Cao, Fei Zhao, Zhaopeng Feng, Zhouhong Gu, Zhenfei Yin, et al. Compbench: Benchmarking complex instruction-guided image editing. *arXiv preprint arXiv:2505.12200*, 2025. 3

[15] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6007–6017, 2023. 1, 3

[16] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 36652–36663. Curran Associates, Inc., 2023. 1

[17] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, et al. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 1, 6

[18] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 36:28541–28564, 2023. 1

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 19730–19742, 2023. 1

[20] Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 3, 4

[21] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision (ECCV)*, pages 366–384, 2024. 5

[22] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, et al. Step1x-edit: A practi-

cal framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 6

[23] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2511–2522, 2023. 1

[24] Yiwei Ma, Jiayi Ji, Ke Ye, Weihuang Lin, Zhibin Wang, Yonghan Zheng, Qiang Zhou, Xiaoshuai Sun, and Rongrong Ji. I2ebench: A comprehensive benchmark for instruction-based image editing. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:41494–41516, 2024. 3

[25] openai. https://openai.com/, 2025. Accessed: 2025-10-13. 1

[26] Yulin Pan, Xiangteng He, Chaojie Mao, Zhen Han, Zeyinzi Jiang, Jingfeng Zhang, and Yu Liu. Ice-bench: A unified and comprehensive benchmark for image creating and editing. *arXiv preprint arXiv:2503.14482*, 2025. 3

[27] Yusu Qian, Jiasen Lu, Tsu-Jui Fu, Xinze Wang, Chen Chen, Yinfei Yang, Wenze Hu, and Zhe Gan. Gie-bench: Towards grounded evaluation for text-guided image editing. *arXiv preprint arXiv:2505.11493*, 2025. 3

[28] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8871–8879, 2024. 1, 2

[29] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. 3, 4, 5

[30] Juntong Wang, Huiyu Duan, Jiarui Wang, Ziheng Jia, Guangtao Zhai, and Xiongkuo Min. Tit-score: Evaluating long-prompt based text-to-image alignment via text-to-image-to-text consistency. *arXiv preprint arXiv:2510.02987*, 2025. 4

[31] Jiarui Wang, Huiyu Duan, Guangtao Zhai, Juntong Wang, and Xiongkuo Min. Aigv-assessor: benchmarking and evaluating the perceptual quality of text-to-video generation with lmm. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 18869–18880, 2025. 4

[32] Jiarui Wang, Huiyu Duan, Yu Zhao, Juntong Wang, Guangtao Zhai, and Xiongkuo Min. Lmm4lmm: Benchmarking and evaluating large-multimodal image generation with lmms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1

[33] Juntong Wang, Jiarui Wang, Huiyu Duan, Guangtao Zhai, and Xiongkuo Min. Tdve-assessor: Benchmarking and evaluating the quality of text-driven video editing with lmms. *arXiv preprint arXiv:2505.19535*, 2025. 1

[34] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18359–18369, 2023. 1, 3

[35] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1

[36] Weicheng Wang, Guoli Jia, Zhongqi Zhang, Liang Lin, and Jufeng Yang. Ps-diffusion: Photorealistic subject-driven image editing with disentangled control and attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 18302–18312, 2025. 4

[37] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13 (4):600–612, 2004. 1

[38] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024. 5

[39] Hongyang Wei, Baixin Xu, Hongbo Liu, Cyrus Wu, Jie Liu, Yi Peng, Peiyu Wang, et al. Skywork unipic 2.0: Building kontext model with online rl for unified multimodal model. *arXiv preprint arXiv:2509.04548*, 2025. 6

[40] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 1, 2, 3, 6, 7, 8

[41] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 1, 6, 7, 8

[42] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *European conference on computer vision (ECCV)*, pages 538–554, 2022. 1

[43] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20144–20154, 2023. 1

[44] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 4

[45] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 1

[46] Bin Xia, Bohao Peng, et al. Dreamomni2: Multimodal instruction-based editing and generation. *arXiv preprint arXiv:2510.06679*, 2025. 1, 2, 6

[47] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the International Conference on Neural Information Processing Systems(NeurIPS)*, pages 15903–15935, 2023. 1

[48] Zitong Xu, Huiyu Duan, Bingnan Liu, Guangji Ma, Jiarui Wang, Liu Yang, Shiqi Gao, Xiaoyu Wang, Jia Wang, Xiongkuo Min, et al. Lmm4edit: Benchmarking and evaluating multimodal image editing with lmms. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACMMM)*, 2025. 1, 3

[49] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025. 4

[50] Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, et al. MiMo-VL Technical Report. *arXiv preprint arXiv:2506.03569*, 2025. 1

[51] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(8):5625–5644, 2024. 3

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3

[53] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8018–8027, 2024. 1

[54] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 1, 6