CompBench: Benchmarking Complex Instruction-guided Image Editing

Bohan Jia[1,*], Wenxuan Huang[1,4,*], Yuntian Tang[1,*], Junbo Qiao[1], Jincheng Liao[1],
Shaosheng Cao[2,✉], Fei Zhao[2], Zhaopeng Feng[5], Zhouhong Gu[6], Zhenfei Yin[7],
Lei Bai[8], Wanli Ouyang[4], Lin Chen[9], Fei Zhao[10], Zihan Wang[1], Yuan Xie[1], Shaohui Lin[1,3,✉,†]
[1]East China Normal University, [2]Xiaohongshu Inc., [3]KLATASDS, MOE, China
[4]The Chinese University of Hong Kong, [5]Zhejiang University, [6]Fudan University, [7]University of Oxford
[8]Shanghai Jiao Tong University, [9]University of Science and Technology of China, [10]Nanjing University
51275901134@stu.ecnu.edu.cn
[*]Equal contribution [✉]Corresponding author [†]Project leader

Abstract

While real-world applications increasingly demand intricate scene manipulation, existing instruction-guided image editing benchmarks often oversimplify task complexity and lack comprehensive, fine-grained instructions. To bridge this gap, we introduce ***CompBench****, a large-scale benchmark specifically designed for complex instruction-guided image editing. CompBench features challenging editing scenarios that incorporate fine-grained instruction following, spatial and contextual reasoning, thereby enabling comprehensive evaluation of image editing models' precise manipulation capabilities. To construct CompBench, we propose an MLLM-human collaborative framework with tailored task pipelines. Furthermore, we propose an instruction decoupling strategy that disentangles editing intents into four key dimensions: location, appearance, dynamics, and objects, ensuring closer alignment between instructions and complex editing requirements. Extensive evaluations reveal that CompBench exposes fundamental limitations of current image editing models and provides critical insights for the development of next-generation instruction-guided image editing systems. Our project page is available at* https://comp-bench.github.io/.

1. Introduction

Recent advances in instruction-guided image editing have pursued user-friendly and efficient manipulation of visual content. While such systems aim to simplify complex editing workflows, real-world applications often demand intricate instructions including spatial relationships, appearance details, and implicit reasoning. This necessitates the development of models with comprehensive capabilities in visual grounding, contextual understanding, and complex reasoning, thereby presenting substantial challenges to existing methodologies. However, as demonstrated in Figure 2, existing instruction-guided image editing benchmarks [14, 30, 44] exhibit critical limitations in assessing these essential capabilities, primarily in three aspects:

Lack of Scene Complexity. A key limitation of current benchmarks is their insufficient scene complexity, which hampers the representation of intricate visual structures inherent in real-world images. Specifically, recent benchmark constructions [22, 42] predominantly source their images from general-purpose datasets, such as MS COCO [19]. While these benchmarks utilize real images, they often present oversimplified, object-centric scenarios with elementary compositions. These scenes typically feature sparse spatial layouts, limited foreground object diversity, and minimal occlusions, lacking the dense object interactions and natural clutter essential for evaluating practical editing capabilities. Consequently, they remain insufficient for comprehensively evaluating models on complex spatial relationships and interactions among multiple objects.

This problem is further exacerbated by benchmark design choices, wherein creators often deliberately exclude highly complex scenes featuring heavy occlusions, intricate details, or dynamic elements due to the challenges they pose for ground truth construction. While this practice facilitates more controllable evaluation, it creates a concerning discrepancy between benchmark performance and real-world applicability.

Consequently, image editing models may attain high metric scores on these relatively simplified benchmarks, yet remain inadequate for real-world editing tasks that demand advanced scene understanding and manipulation. For in-

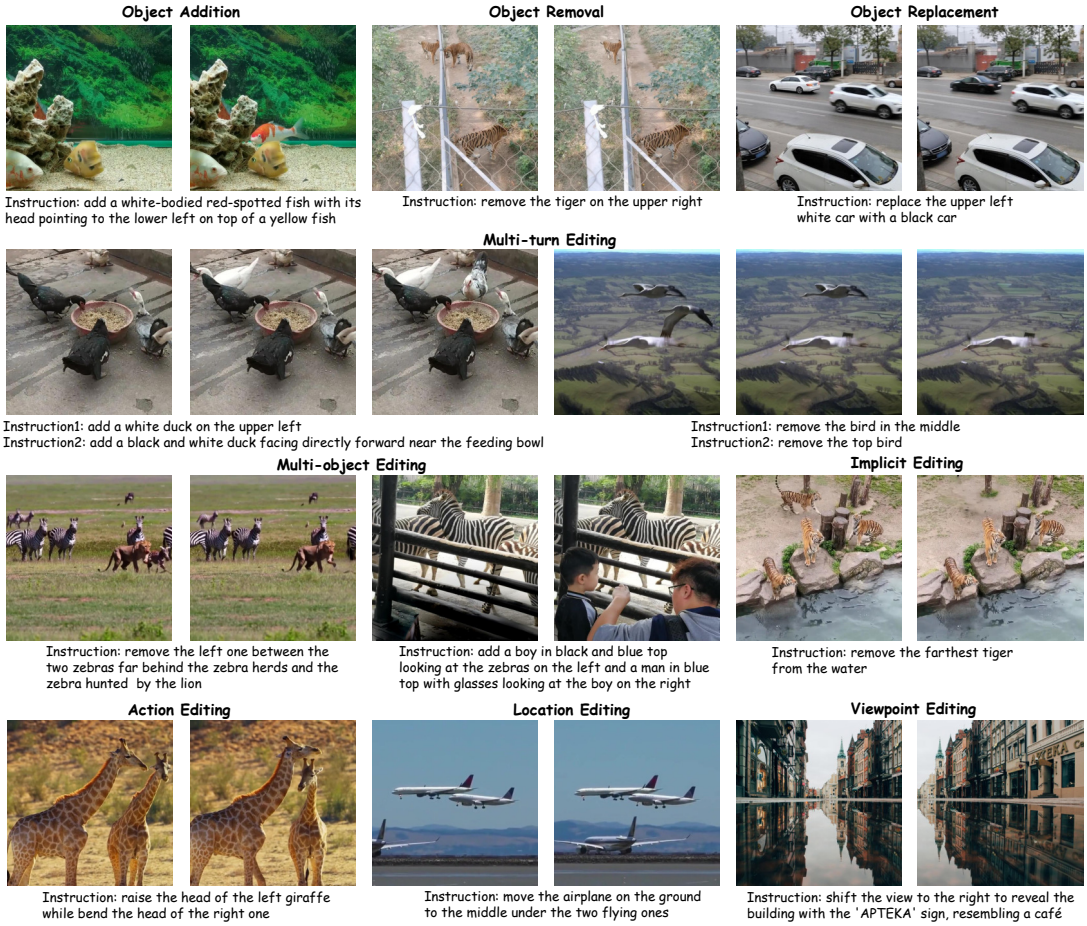


Figure 1. **Examples of CompBench.** The figure showcases nine tasks in our CompBench: object addition, object removal, object replacement, multi-object editing, multi-turn editing, implicit reasoning, action editing, location editing and viewpoint editing.

stance, in reasoning-based tasks, InstructPix2pix [3] exhibits a notable performance decline on our CompBench compared with ReasonEdit [14], showing decreases of approximately 2.5 in PSNR, 0.02 in SSIM, and 0.4 in CLIP-Score.

Limited Instruction and Task Comprehensiveness. Beyond their oversimplified visual scenes, current benchmarks are further constrained by the narrow scope of editing instructions and tasks, failing to reflect the complexity of real-world user demands. Most existing datasets rely on simplistic, atomic-level instructions (*e.g.*, "change the dog to a cat") that lack contextual reasoning, and compositional logic typical of real user requests. In reality, user instructions often require complex reasoning and manipulation. These include multi-object editing ("remove the dog and the cat"), edits based on spatial relationships ("add a man to the right of the woman"), or action editing that modifies dynamic states ("make the man in white bend down more"). Current benchmarks, however, largely neglect these sophisticated task categories. This deficiency in instruction and task diversity prevents models from being rigorously tested on the full spectrum of challenges encountered in real-world applications. Consequently, their performance can be artificially inflated on simple tasks, providing an incomplete and misleading evaluation of true robustness and practical applicability.

Deficiencies in Edited Image Quality. Another critical limitation of current benchmarks is the suboptimal quality of their edited images. Many existing datasets exhibit two predominant issues that compromise their reliability: (1) instruction-alignment inaccuracies, where the edited output fails to precisely fulfill the specified modifications. (2) conspicuous visual artifacts, such as geometric distortions, background inconsistencies, or semantically incoherent objects. These quality deficiencies introduce substantial noise into performance evaluations, potentially leading to misleading assessments of model capabilities. Consequently, such benchmarks may fail to effectively discriminate between truly sophisticated editing systems and those that merely produce superficially plausible but flawed results.

To address the aforementioned issues, we introduce ***CompBench***, the first large-scale benchmark for instruction-guided image editing in complex scenarios, specific examples are illustrated in Figure 1. Our benchmark offers the

Table 1. **Comparison of existing image-editing datasets and benchmarks.** Our benchmark supports seven core editing tasks, including multi-object, action and viewpoint editing, which are absent from most prior benchmarks. Scenario complexity is quantified by four indicators: *Avg. Obj.* (average number of objects per image), *Avg. Cat.* (average number of object categories per image), *OCC* (percentage of images that contain occluded objects), and *OOF* (percentage of images that contain out-of-frame objects). Across all four metrics, our benchmark exhibits the highest complexity, underscoring its suitability for rigorous evaluation.

Datasets / Benchmarks	Size	Types	Task							Complexity			
			Local	Multi-turn	Multi-obj.	Implicit	Action	Location	Viewpoint	Avg. Obj.	Avg. Cat.	Occ. Rate	OOF. Rate
Datasets													
InstructPix2pix [3]	313K	4	✓	✗	✗	✗	✗	✗	✗	8.71	4.16	79.36	81.39
EditWorld [43]	8.6K	1	✗	✗	✗	✓	✗	✗	✗	8.01	4.45	76.67	72.00
UltraEdit [48]	4M	9	✓	✗	✗	✓	✗	✗	✗	7.68	4.70	75.30	78.10
SEED-Data-Edit [12]	3.7M	6	✓	✓	✓	✗	✗	✗	✗	6.21	3.82	63.82	81.40
HQ-Edit [15]	197K	6	✓	✗	✗	✗	✗	✗	✗	8.22	4.84	66.97	60.30
AnyEdit [42]	2.5M	25	✓	✗	✗	✓	✓	✓	✓	6.95	4.37	60.45	57.20
ImgEdit [41]	1.2M	13	✓	✗	✗	✗	✓	✗	✗	9.01	4.72	69.65	69.14
Benchmarks													
MagicBrush [44]	10K	5	✓	✓	✗	✗	✓	✗	✗	9.22	5.04	91.71	78.30
EMU_Edit [30]	–	8	✓	✗	✗	✗	✗	✓	✗	8.38	5.19	78.51	83.60
Reason-Edit [14]	0.2K	-	✓	✗	✗	✓	✗	✗	✗	4.93	3.09	54.30	52.28
I^2EBench [22]	2K	16	✓	✗	✗	✗	✗	✗	✗	7.03	4.20	68.78	66.40
GEdit-Bench [21]	0.6K	11	✓	✗	✗	✗	✓	✗	✗	9.96	4.93	67.67	65.40
Complex-Edit [40]	1K	24	✓	✗	✗	✓	✗	✗	✓	9.23	4.77	78.29	72.98
RefEdit [24]	20K	5	✓	✗	✓	✗	✗	✗	✗	9.74	5.26	91.02	69.00
KRIS-Bench [39]	1.3K	22	✓	✗	✓	✓	✗	✓	✓	6.04	3.09	29.49	40.69
ComplexBench-Edit [34]	763	10	✓	✗	✓	✓	✓	✗	✗	7.85	4.75	75.71	74.14
Ours*	3K	9	✓	✓	✓	✓	✓	✓	✓	**13.58**	**5.87**	**98.47**	**86.38**

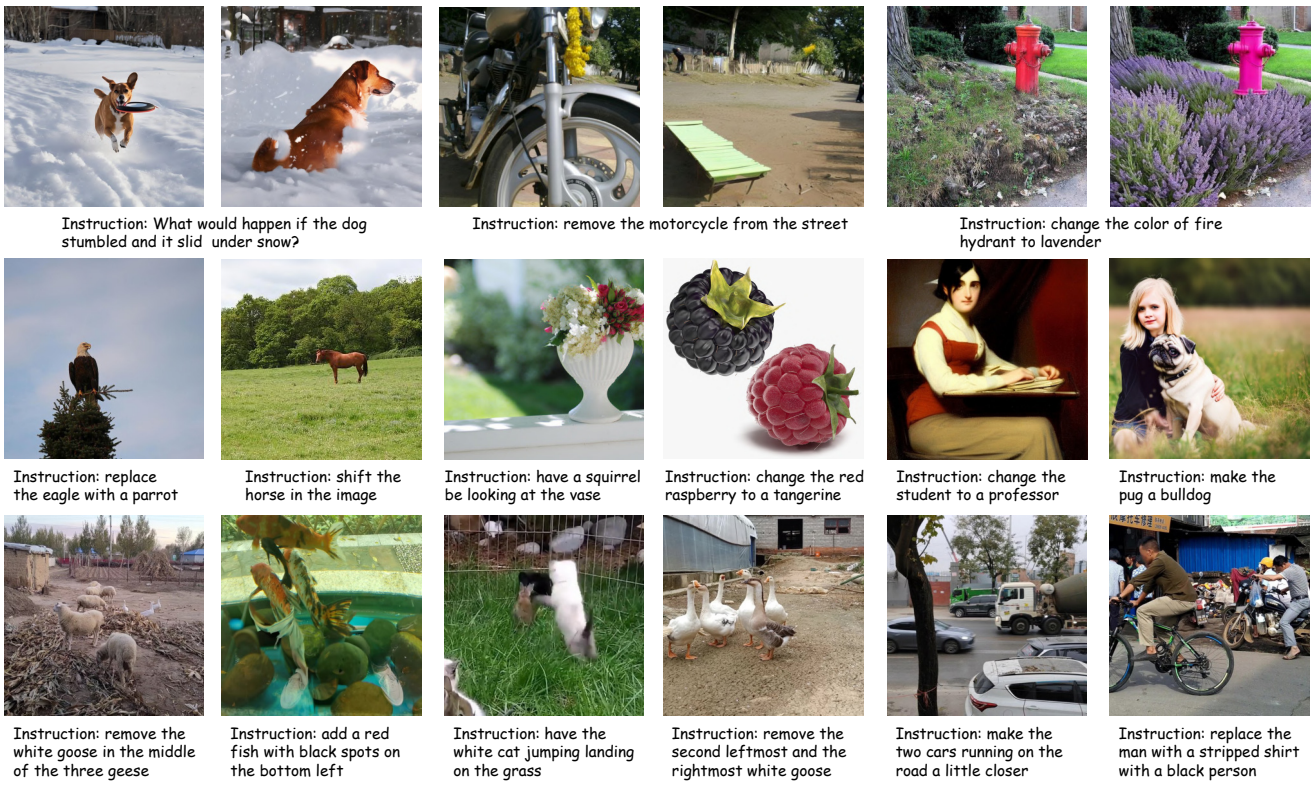


Figure 2. **Comparison between current datasets or benchmarks and our CompBench. First row:** failed cases of other benchmarks. These results fail to maintain background consistencies or introduce noticeable artifacts into the editing region. **Second row:** Examples of other benchmarks. These cases lack scene complexity and instruction comprehensiveness. **Third row:** Examples of our CompBench. Our benchmark features complex real-world scenarios with precise instructions.

following three major advantages:

Realistic and Complex Scene Composition. As shown in Table 1, our benchmark encompasses scenes that embody

the diverse and realistic complexities present in real-world settings. We compare CompBench with existing datasets and benchmarks across four dimensions: average number of objects, average number of object categories, overall object occlusion rate, and out-of-frame object rate. Details of these metrics are shown in the supplementary material. CompBench consistently surpasses prior benchmarks in all these metrics. Notably, our average number of objects per image is approximately **36.3%** higher than the second best (GEdit-Bench [21]), demonstrating the heightened complexity and diversity of our scenes.

Comprehensive Task Coverage and High Difficulty Level. As depicted in Figure 4, CompBench encompasses five major categories, consisting of local editing, multi-editing, action editing, scene spatial editing, and complex reasoning, spanning a total of nine tasks. These tasks are designed to challenge six core capabilities, with a detailed analysis of our benchmark's difficulty for each provided in the supplementary material. Additionally, we propose an Instruction Decomposition Strategy to improve the clarity and precision of image editing instructions. Specifically, we structure editing instructions along four dimensions: spatial positioning (*e.g.*, "left of the table"), visual attributes (such as color or texture), motion states (*e.g.*, "flying"), and object entities. This structured approach converts potentially ambiguous requests into well-defined specifications without sacrificing the natural expressiveness of instructions. By systematically covering each aspect of an editing operation while preserving the flexibility of natural language, our method produces instructions that are both intuitively understandable and technically precise for complex image editing tasks.

High-Quality Data Curation. Every sample in CompBench is meticulously constructed through multiple rounds of expert review, ensuring the highest quality of edits. Unlike other benchmarks where editing failures are common, all data in CompBench represent successfully executed editing results, with SSIM (Structural Similarity Index Measure) scores significantly outperforming those of other datasets, as illustrated in Figure 5. This rigorous quality control ensures that CompBench provides a reliable assessment of model performance in realistically complex editing scenarios.

2. Related Works

Instruction-guided Image Editing. Instruction-guided image editing enables efficient image manipulation using only textual editing instructions, eliminating the need for manual mask or explicit visual inputs and better aligning with user intent. Diffusion models [13], particularly Stable Diffusion [28] (SD), facilitate this task significantly by supporting explicit text inputs. Methods built upon diffusion models such as InstructPix2pix [3], have greatly improved editing effectiveness. InstructPix2pix leverages large language models (LLMs) [4, 6, 32, 33] and text-to-image (T2I) [26–29] models to generate large-scale datasets and trains a diffusion model that is capable of following natural language instructions. HIVE [47] introduces a reward model that leverages human feedback to align edits with human preferences. Approaches such as SmartEdit [14], MGIE [11], and Step1X-Edit [21] integrate image and instruction representations using multi-modal large language models (MLLMs) [1, 18, 20, 35], injecting these capabilities into diffusion models for more precise control. AnyEdit [42] constructs an extremely large-scale multi-task dataset and adopts a mixture-of-experts (MoE) [8, 10] architecture to better accommodate diverse editing tasks. SEED-X [12] utilizes a visual tokenizer to unify image comprehension and generation, establishing a unified multi-granularity comprehension and generation model that enhances editing performance. GoT [9] incorporates Generation Chain-of-Thought [37] reasoning into the editing process, allowing for more refined, step-by-step edits. Recently, FLUX.1 Kontext [16] applies flow matching to build a unified image generation and editing model. Bagel [5] adopts a decoder only architecture to construct a multimodal understanding and generation model. Qwen-Image-Edit [38], the editing model of Qwen-Image [38], demonstrates strong text rendering and image editing capabilities.

Image Editing Benchmarks. High-quality image editing datasets and benchmarks are crucial for model training and evaluation. Several notable benchmarks have been proposed: MagicBrush [44] provides a manually curated 10K dataset covering single-turn, multi-turn, mask-provided, and mask-free editing tasks. EMU-edit [30] introduces a challenging benchmark comprising seven diverse editing tasks. HQ-Edit [15] employs a scalable data collection pipeline to create a high-quality dataset of 200K instruction-guided image editing samples. SmartEdit [14] introduces Reason-Edit, a small-scale, manually curated benchmark focused on complex instruction-based image editing. Edit-world [43] presents the concept of world-instructed image editing and creates a dataset featuring instructions in a world context. I2EBench [22] proposes a comprehensive evaluation benchmark with automated multi-dimensional assessment. UltraEdit [48] develops a scalable framework for producing large and high-quality image editing datasets, introducing a large-scale instruction-based dataset. SEED-Data-Edit [12] provides a hybrid dataset composed of auto-generated, real-world, and human-annotated multi-turn editing samples. More recently, ImgEdit [41] introduces a large scale image editing dataset and a benchmark with multiple aspects. Step1X-Edit [21] construct GEdit-Bench [21] featuring real-world user instructions. Complex-Edit [40] adopts a "Chain-

of-Edit" pipeline to develop an image editing benchmark across instructions of different complexity. ComplexBench-Edit [34] addresses combinatorial reasoning challenges by introducing a benchmark for chain-dependent instructions and a region-specific consistency metric. KRIS-Bench [39] evaluates the cognitive capabilities of models through a knowledge-based taxonomy spanning factual, conceptual, and procedural dimensions. Furthermore, RefEdit [24] targets the precise editing of specific objects in complex, multi-entity scenes based on referring expressions.

3. CompBench

3.1. Task Categorization and Definitions

Our complex instruction-guided image editing benchmark, **CompBench**, contains over 3k image-instruction pairs. For comprehensive evaluation, we categorize the tasks into five major classes encompassing nine specific tasks: (1) Local Editing: manipulating objects via removal, addition, or replacement. (2) Multi-editing: addressing interactions across multi-turn or multi-object editing. (3) Action Editing: modifying dynamic states or object interactions. (4) Scene Spatial Editing: altering spatial properties through location or viewpoint editing. (5) Complex Reasoning: performing implicit contextual edits that require logical reasoning. Examples are illustrated in Figure 1.

3.2. Dataset Generation

In this section, we detailedly demonstrate the generation process of our CompBench(Figure 3).

Source Data Collection and Preprocessing. We utilize the MOSE dataset [7] to address the lack of complex editing data. Image quality is first guaranteed by filtering corrupted frames via automated metrics (*e.g.*, NIQE [45]) and subsequent manual verification. For masks, we decompose multi-object annotations into single-object instances. Finally, discontinuous or heavily occluded masks are discarded using MLLMs (*e.g.*, Qwen-VL [35]), followed by manual refinement to ensure pixel-level precision.

Task-specific Data Generation Pipelines. We design four specialized pipelines to cover diverse complex instruction-guided image editing scenarios: (1) local editing pipeline for object-level manipulations (object removal, object addition, object replacement). (2) action/scene spatial editing pipeline for modifying object dynamics or scene perspectives (action editing, location editing, viewpoint editing). (3) complex reasoning pipeline for implicit contextual edits requiring reasoning (implicit reasoning). (4) multi-editing pipeline for multi-object and multi-turn editing tasks. All pipelines adopt a unified MLLM-Human Collaborative Framework: multimodal large language models (MLLMs) [1, 18, 20, 35] generate initial task-specific instructions by analyzing visual scenes and editing goals, followed by human validation to ensure instruction-image alignment and image editing fidelity. Unsuccessful edits are iteratively re-generated or discarded, retaining only high-fidelity samples. Detailed procedures are provided in the supplementary material.

Instruction Decomposition Strategy. To enhance the clarity and precision of editing instructions, we propose a structured framework that organizes editing instructions along four aspects: spatial positioning, visual attributes, motion states, and object entities. This approach transforms ambiguous editing requests into well-defined specifications while maintaining natural expressiveness. The method employs a two-phase generation process: first, an MLLM produces dimension-aware instruction candidates by analyzing visual contexts. Then human experts refine these to ensure precision and consistency. By systematically addressing each aspect of the editing operation while preserving the flexibility of natural language, this framework enables the creation of instructions that are both intuitively understandable and technically precise for complex image editing tasks.

Characteristics and Statistics. As illustrated in Figure 4, our benchmark comprises five major categories encompassing nine complex editing tasks, yielding over 3000 high-quality samples. We employ SSIM [36] to evaluate semantic consistency between pre- and post-edited images. As shown in Figure 5, CompBench achieves notably higher SSIM than other datasets and benchmarks. Notably, our dataset features significantly more challenging editing tasks requiring comprehensive capabilities such as visual grounding and complex reasoning. Quantitative indicators (*e.g.*, average number of objects and categories) also demonstrate that our benchmark features substantially higher scene complexity. Detailed subtask definitions and core competency analyses are provided in the supplementary material.

4. Experiments

4.1. Settings

Baselines. Our study focuses on instruction-guided image editing models, excluding those based on global description guidance. The evaluated models include: InstructPix2pix [3], MagicBrush [44], HIVE [47], Smartedit [14], MGIE [11], HQ-Edit [15], CosXL-Edit [31], UltraEdit [48], AnyEdit [42], Seed-X-Edit [12], GoT [9], Step1X-Edit [21], Bagel [5], FLUX.1 Kontext [16], and Qwen-Image-Edit [38].

Evaluation Metrics and Methods. For local editing, multi-editing, and implicit reasoning tasks, we employ a foreground-background decoupling strategy. To evaluate

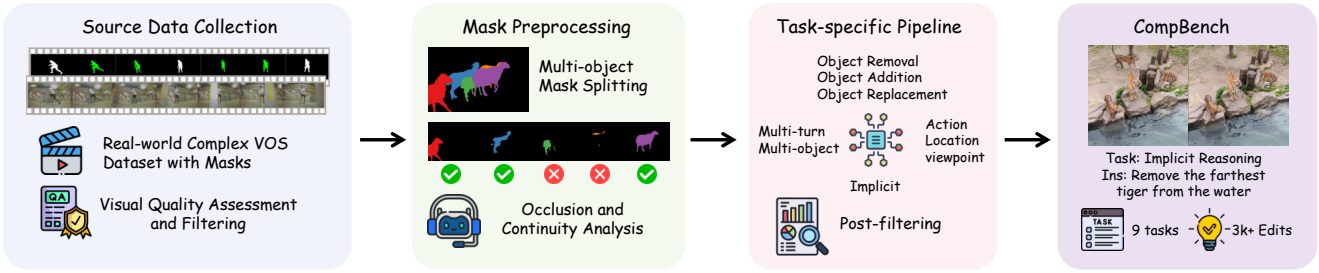


Figure 3. **The construction pipeline of CompBench.** The pipeline consists of two main stages: (a) Source data collection and preprocessing, wherein high-quality data are identified through image quality filtering, mask decomposition, occlusion and continuity evaluation, followed by thorough human verification. (b) Task-specific data generation using four specialized pipelines within our MLLM-Human Collaborative Framework, where multimodal large language models generate initial editing instructions that are subsequently validated by humans to ensure high-fidelity, semantically aligned instruction-image pairs for complex editing tasks.

Table 2. **Evaluation results on local editing, multi-object editing and implicit reasoning.** LC-T denotes local CLIP scores between the edited foreground and the local description. LC-I refers to the CLIP image similarity between the foreground edited result and ground truth (GT) image. Top-three evaluation results are highlighted in **red** (1st), **blue**(2nd), and **green** (3rd).

Model	Local Editing					Multi-object Editing					Implicit Reasoning				
	Foreground		Background			Foreground		Background			Foreground		Background		
	LC-T ↑	LC-I ↑	PSNR(dB) ↑	SSIM ↑	LPIPS ↓	LC-T ↑	LC-I ↑	PSNR(dB) ↑	SSIM ↑	LPIPS ↓	LC-T ↑	LC-I ↑	PSNR(dB) ↑	SSIM ↑	LPIPS ↓
InstructPix2pix [3]	19.269	0.778	21.828	0.706	0.124	20.050	0.804	20.534	0.671	0.152	18.981	0.794	21.813	0.683	0.125
MagicBrush [44]	20.051	0.798	23.429	0.741	0.090	20.004	0.821	24.176	0.738	0.081	19.498	0.826	22.143	0.714	0.106
HIVE-w [47]	19.804	0.771	19.903	0.641	0.198	19.949	0.782	19.896	0.624	0.210	18.590	0.777	20.261	0.602	0.219
HIVE-c [47]	19.226	0.773	21.732	0.689	0.147	19.904	0.798	21.854	0.676	0.144	18.888	0.787	22.167	0.666	0.144
Smart-edit-7B [14]	19.999	0.799	24.389	0.761	0.074	20.186	0.825	**24.886**	0.745	**0.076**	**19.740**	0.831	23.060	0.732	0.096
MGIE [11]	18.773	0.764	18.204	0.639	0.257	20.102	0.742	15.380	0.500	0.354	18.304	0.795	22.144	0.719	0.151
CosXL-Edit [31]	19.068	0.778	20.809	0.712	0.148	19.606	0.807	21.068	0.698	0.153	18.003	0.800	21.190	0.683	0.156
HQ-Edit [15]	18.888	0.734	11.768	0.411	0.428	19.401	0.708	13.032	0.432	0.400	18.688	0.766	11.877	0.387	0.459
UltraEdit [48]	19.690	0.787	22.932	0.741	0.143	20.081	0.810	22.725	0.732	0.152	18.289	0.786	23.362	0.717	0.142
AnyEdit [42]	19.994	0.797	22.812	0.716	0.124	20.257	0.810	23.493	0.710	0.116	19.572	0.816	20.276	0.639	0.191
SEED-X [12]	17.900	0.780	21.456	0.805	0.139	19.418	0.835	21.166	0.798	0.148	17.437	0.782	21.438	0.790	0.134
GoT [9]	20.268	0.807	24.675	0.889	**0.067**	20.225	0.827	22.486	0.842	0.108	19.246	0.821	**24.889**	0.860	0.088
Step1X-Edit [21]	20.495	0.817	23.372	0.882	0.078	20.459	0.861	23.782	**0.886**	0.078	19.175	0.847	23.408	**0.869**	**0.083**
Bagel [5]	**21.059**	**0.838**	**27.692**	**0.935**	**0.045**	**20.856**	**0.883**	**26.849**	**0.932**	**0.051**	**19.719**	**0.874**	**28.891**	**0.919**	**0.052**
FLUX.1 Kontext [16]	**21.328**	**0.821**	**25.613**	**0.941**	**0.050**	**21.020**	**0.868**	**26.278**	**0.952**	**0.048**	19.596	**0.867**	**25.401**	**0.932**	**0.061**
Qwen-Image-Edit [38]	**21.522**	**0.828**	**24.968**	**0.892**	0.072	**20.722**	**0.864**	23.492	0.826	0.103	**20.046**	**0.859**	22.815	0.775	0.124

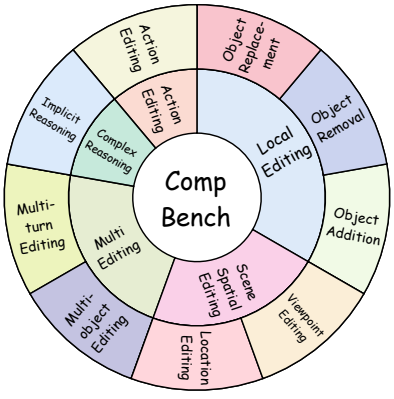


Figure 4. **Task taxonomy of CompBench.** Illustration of the full range of task types in CompBench.

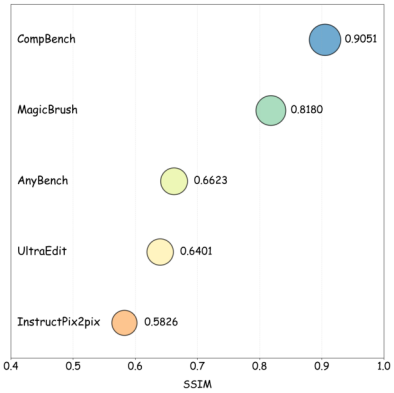


Figure 5. **SSIM comparison among different datasets and benchmarks.** Note that UltraEdit [48] and InstructPix2pix [3] are datasets, whereas the remaining entries are benchmarks.

background preservation, we compute PSNR, SSIM [36], and LPIPS [46] on the background regions. For foreground evaluation, we assess two aspects using CLIP [25]: *editing accuracy* via CLIP image similarity between the localized edited foreground and the ground truth (GT) image (denoted as LC-I), and *instruction following* via local CLIP scores between the edited foreground and the local description (denoted as LC-T).

Table 3. **Evaluation results on multi-turn editing.**

Model	Turn1					Turn2				
	Foreground		Background			Foreground		Background		
	LC-T	LC-I	PSNR	SSIM	LPIPS	LC-T	LC-I	PSNR	SSIM	LPIPS
InstructPix2pix [3]	19.617	0.786	21.361	0.682	0.134	19.856	0.777	17.764	0.573	0.233
MagicBrush [44]	19.564	0.811	24.099	0.731	0.089	19.694	0.809	21.294	0.683	0.133
HIVE-w [47]	20.012	0.781	20.065	0.622	0.195	**20.389**	0.773	17.315	0.533	0.270
HIVE-c [47]	19.922	0.789	21.331	0.660	0.154	19.964	0.779	18.350	0.590	0.217
Smart-edit-7B [14]	19.595	0.813	24.653	0.740	0.080	19.571	0.808	**23.383**	0.723	**0.104**
MGIE [11]	19.194	0.814	18.473	0.602	0.246	19.328	0.803	14.820	0.493	0.366
HQ-Edit [15]	19.364	0.758	12.384	0.391	0.410	19.228	0.753	11.608	0.351	0.478
CosXL-Edit [31]	19.670	0.790	20.292	0.681	0.167	19.430	0.769	16.390	0.566	0.312
UltraEdit [48]	19.816	0.793	23.065	0.737	0.151	20.015	0.793	19.927	0.666	0.208
AnyEdit [42]	19.858	0.812	23.398	0.711	0.113	20.061	0.806	20.001	0.633	0.189
SEED-X [12]	19.576	0.796	21.049	0.792	0.153	19.188	0.769	17.708	0.629	0.280
GoT [9]	19.833	0.815	24.840	**0.891**	**0.067**	19.832	0.809	**23.487**	**0.855**	0.107
Step1X-Edit [21]	19.873	0.833	23.802	0.888	0.078	19.754	0.834	21.016	0.832	0.125
Bagel [5]	**20.060**	**0.852**	**28.959**	**0.949**	**0.036**	20.007	**0.850**	**23.941**	**0.897**	**0.085**
FLUX.1 Kontext [16]	**20.634**	**0.845**	**26.193**	**0.954**	**0.046**	**20.620**	**0.840**	22.417	**0.907**	**0.094**
Qwen-Image-Edit [38]	**20.468**	**0.845**	**24.863**	0.846	0.088	**20.584**	**0.838**	20.814	0.783	0.151

Additionally, for action, location, and viewpoint editing tasks—where the object's morphology, position, or viewpoint may change significantly—automatic metrics alone are insufficient. To address this, we introduce multi-perspective scoring via GPT-4o [23], Qwen2.5-VL-72B [2], and human annotators. We design tailored prompts instructing the models to rate editing performance on a 0-10 scale. Simultaneously, trained human annotators evaluate background fidelity, editing intent, instruction following, and artifact presence based on standardized guidelines. Detailed prompts, annotation instructions, and additional evaluation results are available in the supplementary material.

4.2. Experiment Results

The experimental results for local editing, multi-turn editing, multi-object editing, implicit reasoning, and action/location/viewpoint editing are presented in Tables 2, 3, and 4, respectively. Our key analysis of the results are as follows: (1) No model dominates across all tasks. Among all evaluated models, Bagel [5] emerges as the most prominent one, achieving top results in 18 out of 37 metrics (nearly 50%) across 9 tasks. Notably, Bagel [5], Qwen-Image-Edit [38], and FLUX.1 Kontext [16] consistently deliver superior performance, securing top-three rankings in the majority of metrics across most tasks, followed by Step1X-Edit [21]. In contrast, HQ-Edit [15] demonstrates substantially inferior results in nearly all tasks. (2) For multi-turn editing tasks, all models exhibit a notable decline in background consistency metrics during the second editing round. Among them, SmartEdit [14] maintains relatively robust performance in the second editing turn. (3) Bagel [5] consistently leads in the LC-I metric, reflecting its superior foreground fidelity. Additionally, it ranks high in background consistency, effectively preserving spatial and contextual information during editing. (4) For the more challenging action/location/viewpoint editing tasks, Qwen-Image-Edit [38] and Bagel [5] perform comparably and significantly outperform most other models. Step1X-Edit [21] also exhibits promising editing performance in these scenarios.

5. Insights

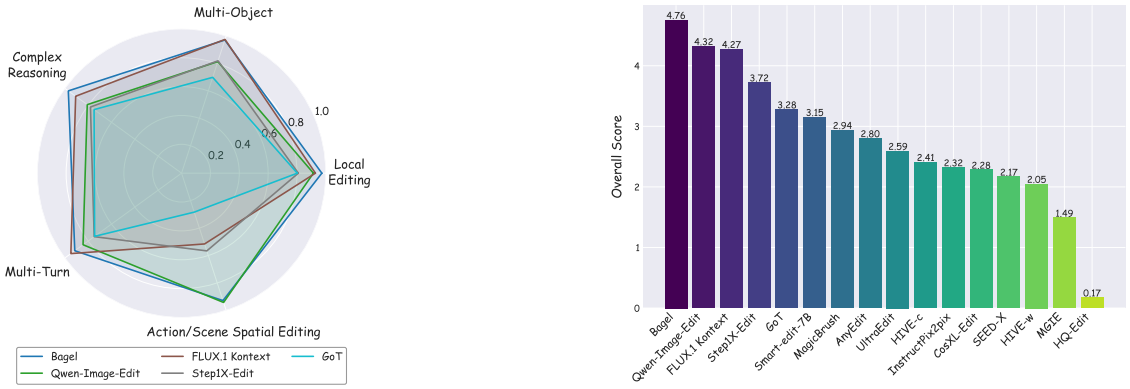
In this section, we provide deeper insights and future directions based on model performances and failure analysis on our proposed CompBench.

The Critical Role of MLLMs. We observe a strong correlation between architectural design and editing performance. As shown in Figure 6(a) and (b) (calculation details in supplementary), the top-performing models (Bagel [5], Qwen-Image-Edit [38], FLUX.1 Kontext [16], and Step1X-Edit [21]) mostly integrate multi-modal large language models (MLLMs). Models lacking MLLM integration frequently ignore instructions or edit wrong targets, indicating that standard CLIP-text alignment is insufficient for complex reasoning. Therefore, integrating MLLMs is indispensable for accurately interpreting intricate textual instructions and visual contexts.

Planner-Executor Misalignment. Despite the advantages of MLLMs, our failure analysis reveals a widespread planner-executor misalignment. In high-density scenes, even when MLLMs (Planner) correctly identify targets, diffusion models (Executor) often fail at precise masking, causing background leakage. Unified architectures help mitigate this mismatch, as exemplified by Bagel's joint learning of understanding and generation. Consequently, future research should focus on improving pixel-level grounding stability in cluttered scenes, moving beyond purely semantic understanding.

Table 4. **Comparison on Action, Location, and Viewpoint Editing.** Results for GPT-4o, Qwen-72B, Human Evaluation, and Average scores (top-3 per column highlighted in red, blue, green).

Model	Action				Location				Viewpoint			
	GPT	Qwen	Human	Avg.	GPT	Qwen	Human	Avg.	GPT	Qwen	Human	Avg.
InstructPix2pix [3]	3.047	1.124	3.101	2.424	3.425	2.167	2.859	2.817	0.699	0.482	0.036	0.406
MagicBrush [44]	3.511	1.449	3.584	2.848	4.603	2.260	3.717	3.527	0.892	0.410	0.108	0.470
HIVE-w [47]	3.151	1.764	3.067	2.661	4.110	2.192	3.421	3.241	1.494	0.283	0.036	0.604
HIVE-c [47]	3.977	1.596	3.797	3.123	4.192	2.470	3.558	3.407	2.193	0.675	0.145	1.004
Smart-edit-7B [14]	4.233	1.607	4.348	3.396	3.890	2.875	3.505	3.423	2.169	0.590	0.410	1.056
MGIE [11]	1.921	1.213	1.797	1.644	1.726	1.795	1.728	1.750	0.205	0.193	0	0.133
CosXL-Edit [31]	4.270	2.375	3.966	3.537	**5.479**	2.493	4.517	4.163	1.916	0.988	0.301	1.068
HQ-Edit [15]	1.449	0.528	1.033	1.003	1.425	0.726	1.079	1.077	0.470	0.289	0	0.253
UltraEdit [48]	4.449	1.807	4.235	3.497	4.014	2.055	3.410	3.160	0.494	0.706	0	0.400
AnyEdit [42]	3.750	0.978	3.168	2.632	5.068	2.479	4.178	3.908	1.687	0.783	0.072	0.847
SEED-X [12]	2.270	1.494	1.685	1.816	3.028	3.247	2.771	3.015	2.241	1.169	0	1.137
GoT [9]	3.337	1.989	3.134	2.820	3.625	3.192	3.164	3.327	0.916	0.675	0.446	0.679
Step1X-Edit [21]	**6.270**	**3.944**	**5.348**	**5.187**	5.041	**4.479**	**4.786**	**4.769**	2.470	1.205	0.663	1.446
Bagel [5]	**6.899**	**5.056**	**6.629**	**6.195**	**7.137**	**6.233**	**6.219**	**6.530**	**5.193**	**3.892**	**4.663**	**4.583**
FLUX.1 Kontext [16]	5.169	3.202	4.517	4.296	3.000	3.110	3.836	3.315	**3.471**	**2.373**	**3.108**	**2.984**
Qwen-Image-Edit [38]	**6.910**	**5.382**	**6.764**	**6.352**	**7.055**	**5.096**	**4.658**	**5.603**	**6.193**	**4.470**	**6.181**	**5.615**



(a) (b)

Figure 6. **Overall Model Performance.**(a) Top 5 model performance in five major tasks. (b) Overall model performance across all tasks.

Reasoning Bottleneck. Beyond basic semantic alignment, complex multi-modal reasoning is foundational for high-fidelity edits. Enhancing reasoning capabilities significantly boosts performance, evident in data-centric strategies (e.g., SmartEdit training on LISA [17]'s reasoning segmentation data) and method-centric designs (e.g., GoT introducing Chain-of-Thought [37] via MLLMs). Therefore, a crucial direction for future work involves continuously enhancing the reasoning capabilities of MLLMs through reasoning-aware training paradigms to ensure precise intent interpretation.

Geometric Hallucination. Beyond semantics, we observe severe limitations in physical consistency. For complex spatial tasks like Action and Viewpoint editing, models frequently hallucinate distorted geometries. To address this issue, future frameworks need to incorporate 3D structural priors or geometric guidance into 2D-trained editors to maintain strict physical consistency during generation.

6. Conclusion

In this work, we introduce CompBench, the first large-scale benchmark designed for complex instruction-guided image editing. Our benchmark encompasses five major categories with nine specialized tasks comprising over 3,000 high-quality image editing pairs with corresponding instructions. We conduct extensive evaluation to systematically assess the capabilities and limitations of contemporary editing systems and validate our evaluation framework. Our findings not only reveal significant performance gaps in current models but also provide valuable insights to guide future research toward next-generation image editing systems.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (NO. 62572193), the Open Research Fund of the Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Ministry of Education, and the Fundamental Research Funds for the Central Universities.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 4, 5

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 7

[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*, pages 18392–18402, 2023. 2, 3, 4, 5, 6, 7, 8

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 4

[5] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 4, 5, 6, 7, 8

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 4

[7] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20224–20234, 2023. 5

[8] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, pages 5547–5569. PMLR, 2022. 4

[9] Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025. 4, 5, 6, 7, 8

[10] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 4

[11] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023. 4, 5, 6, 7, 8

[12] Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007*, 2024. 3, 4, 5, 6, 7, 8

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4

[14] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8362–8371, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[15] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024. 3, 4, 5, 6, 7, 8

[16] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 4, 5, 6, 7, 8

[17] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9579–9589, 2024. 8

[18] 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 *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 4, 5

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 4, 5

[21] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 3, 4, 5, 6, 7, 8

[22] 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*, 37:41494–41516, 2024. 1, 3, 4

[23] OpenAI. GPT-4o System Card. https://cdn.openai.

com/gpt-4o-system-card.pdf, 2024. Accessed: 2025-05-16. 7

[24] Bimsara Pathiraja, Maitreya Patel, Shivam Singh, Yezhou Yang, and Chitta Baral. Refedit: A benchmark and method for improving instruction-based image editing model on referring expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15646–15656, 2025. 3, 5

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6

[26] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 4

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

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 4

[30] 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*, pages 8871–8879, 2024. 1, 3, 4

[31] Stability AI. Cosxl - stable diffusion model. https://huggingface.co/stabilityai/cosxl, 2024. Accessed: 2025-05-16. 5, 6, 7, 8

[32] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 4

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[34] Chenglin Wang, Yucheng Zhou, Qianning Wang, Zhe Wang, and Kai Zhang. Complexbench-edit: Benchmarking complex instruction-driven image editing via compositional dependencies. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13391–13397, 2025. 3, 5

[35] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 4, 5

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5, 6

[37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 4, 8

[38] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 4, 5, 6, 7, 8

[39] Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. Kris-bench: Benchmarking next-level intelligent image editing models. *arXiv preprint arXiv:2505.16707*, 2025. 3, 5

[40] Siwei Yang, Mude Hui, Bingchen Zhao, Yuyin Zhou, Nataniel Ruiz, and Cihang Xie. Complex-edit: Cot-like instruction generation for complexity-controllable image editing benchmark. *arXiv preprint arXiv:2504.13143*, 2025. 3, 4

[41] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025. 3, 4

[42] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26125–26135, 2025. 1, 3, 4, 5, 6, 7, 8

[43] Bohan Zeng, Ling Yang, Jiaming Liu, Minghao Xu, Yuanxing Zhang, Pengfei Wan, Wentao Zhang, and Shuicheng Yan. Editworld: Simulating world dynamics for instruction-following image editing. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12674–12681, 2025. 3, 4

[44] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023. 1, 3, 4, 5, 6, 7, 8

[45] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 5

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[47] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese,

Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9026–9036, 2024. 4, 5, 6, 7, 8

[48] Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024. 3, 4, 5, 6, 7, 8