BenchCouncil Transactions on Benchmarks, Standards and Evaluations 4 (2024) 100201

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/

Full length article

Patrick Star: A comprehensive benchmark for multi-modal image editing

Di Cheng [a], ZhengXin Yang [b], ChunJie Luo [b], Chen Zheng [c,d,e], YingJie Shi [a,*]

[a] School of Arts & Sciences, Beijing Institute of Fashion Technology, Beijing, China
[b] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[c] Institute of Software, Chinese Academy of Sciences, Beijing, China
[d] University of Chinese Academy of Sciences, Nanjing, China
[e] Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China

ARTICLE INFO

Keywords:
Image editing
AIGC
Benchmark

ABSTRACT

Generative image editing enhances and automates traditional image designing methods. However, there is a significant imbalance in existing research, where the development of sketch-guided and example-guided image editing has not been sufficiently explored compared to text-guided image editing, despite the former being equally important in real-world applications. The leading cause of this phenomenon is the severe lack of corresponding benchmark datasets. To address this issue, this paper proposes a comprehensive and unified benchmark dataset, Patrick Star, which consists of approximately 500 test images, to promote balanced development in this field across multi-task and multi-modal settings. First, theoretical analysis grounded in Evaluatology highlights the importance of establishing a balanced benchmark dataset to advance research in image editing. Building on this theoretical foundation, the dataset's construction methodology is explained in detail, ensuring it addresses critical gaps in existing studies. Next, statistical analyses are conducted to verify the dataset's usability and diversity. Finally, comparative experiments underscore the dataset's potential as a comprehensive benchmark, demonstrating its capacity to support balanced development in image editing.

1. Introduction

Image editing has emerged as a crucial direction in both industry and academia, particularly as digital content creation becomes increasingly central to modern communication, entertainment, and business operations. Modern generative image editing approaches leverage deep learning models that use conditional information as guidance to achieve intelligent image manipulation. These approaches overcome traditional limitations not only by reducing editing time and improving efficiency, but also by lowering the technical barriers for users. Additionally, these AI-powered editing tools have revolutionized the creative workflow by enabling more intuitive and precise control over image modifications, marking a significant departure from conventional pixel-level manipulation methods. The field of image editing has attracted increasing research attention, evolving from single-modal to multi-modal approaches. Various image editing tasks have been developed, including but not limited to text-guided image editing, sketch-guided image editing, and example-based image editing. Each modality offers unique advantages: text guidance provides natural language interaction, sketch guidance enables precise spatial control, and example-guided approaches allow for intuitive style and content transfer. The integration of these different modalities has opened new possibilities for more flexible and powerful image editing systems.

Recent research [1] reveals an imbalance in the development of these three image-editing approaches. The emergence of CLIP [2] sparked significant advances in text–image alignment, leading to a boom in text-guided image editing research. Further more, the widespread adoption of text prompts in commercial applications, due to their user-friendly interaction mode, has inadvertently led to relatively less attention being paid to sketch-guided and example-guided editing approaches. However, alternative guidance methods are equally important as guided approaches in the field of image editing, particularly in scenarios where precise visual control or style matching is crucial. Sketch-guided editing, for instance, offers invaluable advantages in professional design workflows where exact spatial arrangements are required, while example-guided methods excel in maintaining visual consistency and achieving complex shape transfers that may be difficult to describe through text alone.

A fundamental shift in research paradigms lies at the root of this imbalance. As the field evolves from image generation to image editing, the nature of sketch-guided image editing tasks has transcended traditional image translation. However, this transformation appears to

* Corresponding author.
E-mail address: shiyingjie1983@163.com (Y. Shi).

https://doi.org/10.1016/j.tbench.2025.100201
Received 13 February 2025; Received in revised form 29 March 2025; Accepted 8 April 2025
Available online 30 April 2025

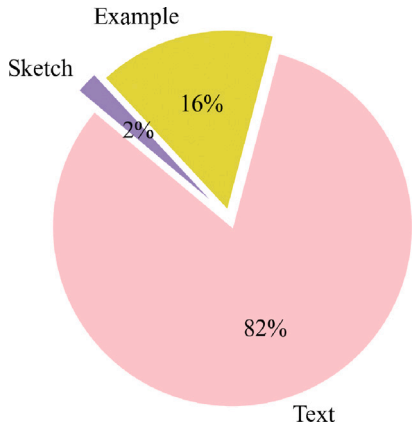

Fig. 1. The distribution of image editing tasks is imbalanced.

have been overlooked by many researchers who may perceive the field as exhausted, failing to recognize the new possibilities and challenges that emerge in the context of editing rather than generation. Meanwhile, example-guided image editing remains at a nascent stage, with current approaches primarily limited to surface-level manipulations and complete object transfers. These methods have yet to achieve the granular control and sophisticated content manipulation capabilities that modern image editing demands. The relative immaturity of these approaches stems from their current inability to handle partial object modifications or more nuanced content transformations. This fundamental misalignment in research focus and the early developmental stage of example-guided methods have contributed to a significant disparity in benchmark datasets compared to text-guided image editing tasks. The lack of comprehensive benchmarks is not merely a data collection issue, but rather a reflection of the deeper challenges in understanding and defining the full potential of these guidance modalities in the context of modern image editing.

Based on our literature review [3], as shown in Fig. 1, current research in the field of image editing exhibits a significant imbalance across different guidance modalities. This uneven distribution of research attention has led to two critical challenges: First, when conducting comparative experiments, niche tasks such as sketch-guided and example-guided editing struggle to find appropriate control groups. Researchers often resort to direct comparisons with text-guided methods, a practice that presents methodological limitations. Second, existing benchmark datasets suffer from two major deficiencies:

- Fragmentation of Test Content: Test data for different guidance modalities (text, sketch, and example) are typically isolated, with almost no benchmark datasets supporting evaluation across all three editing modes simultaneously. This fragmentation makes cross-modal performance comparisons both challenging and less convincing.
- Inconsistency in Evaluation Metrics: Taking Pre_error (used to evaluate the preservation of non-edited regions) as an example, the accuracy of such evaluation metrics heavily depends on the uniformity of editing region sizes. However, existing benchmark datasets often overlook this crucial factor by not standardizing the editing regions, directly impacting the comparability and reliability of evaluation results.

We present Patrick Star Bench, a benchmark dataset designed specifically for image editing tasks. Addressing the issues of inconsistent evaluation standards and fragmented test content in existing datasets, we developed a systematic benchmark construction method from the perspective of scientific evaluation. as illustrated in Fig. 2. This comprehensive dataset consists of five components:Source Image, Mask, Prompt, Sketch, Example and GroundTruth. Patrick Star encompasses 500 sets totaling 2,500 images, covering three major tasks and seven subtasks. The main characteristics and contributions of this benchmark dataset are as follows: First, it achieves unified support for three major editing modalities text-guided, sketch-guided, and example-guided editing, providing a reliable standard for evaluating model performance across cross-modal editing tasks. The dataset implements a rigorous quality control system, ensuring alignment between prompts and images, precision of mask boundaries, and clarity of line extractions. Through experimental validation on six representative models, Patrick Star Bench demonstrates strong discriminative power and reliability. The experimental results indicate that this benchmark not only effectively evaluates performance differences across various editing methods but also provides valuable reference points for subsequent model improvements. we developed a systematic benchmark construction method from the perspective of scientific evaluation. The main contributions are:

(1) Cross-Modal Integration: For the first time, a benchmark dataset unifies the evaluation of text-guided, sketch-guided, and example-guided editing within a single framework. This directly addresses the fragmentation challenge in existing benchmarks and enables reliable cross-modal performance comparisons.

(2) Standardized Evaluation Framework: By implementing consistent mask regions and evaluation metrics across all modalities, our benchmark resolves the long-standing issue of inconsistent evaluation standards, particularly in measuring preservation errors (Pre_error) across different editing approaches.

(3) Extensive Validation: Through rigorous experiments with six representative models, Patrick Star Bench demonstrates strong discriminative power in identifying performance differences across various editing methods.

These contributions collectively address the key challenges in current image editing evaluation and establish a more robust and comprehensive evaluation standard for the field.

2. Related work

2.1. Image editing methods

Image editing has evolved from single-modal approaches to multimodal methodologies, encompassing text-guided, sketch-guided, and example-guided editing techniques. Among these, text-guided image editing has experienced remarkable growth with numerous downstream applications, including instructional editing, position modification [4,5], object manipulation [6–9] (movement, deletion, and addition), and scene reconstruction. These methods typically leverage pre-trained models through fine-tuning or task-specific adapters, achieving impressive results while reducing computational costs. The technical paradigm in image editing has shifted from GANs to Diffusion Models, with each advancement demanding larger datasets and more parameters. However, this trend toward increasingly resource-intensive models poses challenges for general users who may lack access to sufficient computational resources. Sketch-guided image editing has undergone a significant transformation. In its early stages, the field primarily focused on direct image-to-image translation within sketched regions. However, contemporary approaches have evolved to utilize sketches as auxiliary conditional controls for content generation in specific domains. Despite this advancement, the application of sketch guidance in local image editing tasks remains relatively unexplored, representing a notable gap in current research. Example-guided image editing currently encompasses two primary approaches. The first method focuses on object transfer through segmentation, which preserves the complete set of object characteristics but often struggles with seamless integration, particularly when dealing with non-independent objects. The second approach leverages semantic information extracted from reference images to generate semantically consistent results. While this method excels at contextual integration, it may not fully

Fig. 2. Patrick Star: Cases for Image editing tests.

Fig. 3. COCOEE pair of test cases.

preserve specific object details. This creates a fundamental trade-off between feature preservation and contextual harmony, highlighting the challenge of balancing object fidelity with seamless scene integration in example-guided editing.

2.2. Image editing evaluation benchmarks

While numerous evaluation benchmarks exist in the image editing field, these benchmark datasets commonly exhibit significant limitations. As shown in Table 1, existing benchmarks typically support only a single guidance modality: text guidance (e.g., EditBench [10], TedBench [5]), example guidance (e.g., COCOEE [11]), or sketch guidance (e.g., SKETCH Dataset [12]). Through in-depth analysis of existing benchmarks, we identify several key challenges:

First, the limitation of evaluation paradigms. EditBench [10] primarily focuses on text and mask-guided inpainting while neglecting global editing tasks; TedBench [5], despite expanding the task scope, lacks detailed instructions; EditVal [13] is constrained by the low resolution and blurry image quality inherited from the MS-COCO dataset [14] and Emu Edit relies solely on input images from the MagicBrush [15] benchmark. Such singular evaluation perspectives fail to comprehensively reflect model performance.

Second, the absence of cross-modal support. Although COCOEE [11] attempts to support multiple guidance modalities through data processing, its scope remains limited to simple editing tasks within object detection contexts. As illustrated in Fig. 3, this dataset exhibits inconsistencies between reference images and ground truth, highlighting the technical challenges in constructing high-quality multi-modal editing benchmarks: maintaining complex non-independent editing elements (such as modifying a round collar to a notched lapel) while ensuring content and style consistency between target and groundtruth images.

To address these challenges, we propose Patrick Star Bench with a more systematic task classification system. For simple tasks, it includes quantity changes color modifications, position adjustments, and basic state transformations, focusing on evaluating models' local precise editing capabilities. For complex tasks, it encompasses material transformations, content synthesis, overall consistency, and texture transformations, comprehensively testing models' ability to handle sophisticated editing scenarios. This classification not only covers the seven types of editing operations in traditional benchmarks (background modification, global transformation, style transfer, object removal, addition, local editing, and texture/color changes) but also provides more fine-grained evaluation criteria.

More importantly, Patrick Star Bench pioneers unified support for text, example, and sketch guidance, establishing a more comprehensive and reliable standard for evaluating cross-modal image editing capabilities. Through systematic data construction processes and strict quality control, we have successfully addressed the challenges of data consistency and evaluation standard uniformity.

3. Dataset construction

3.1. Semantic tag specification

To standardize prompt generation and validation processes in image editing tasks, we designed a strict semantic tag specification. As show in Fig. 7, this specification consists of six fundamental semantic tags: Position tags <P>, Object tags <O>, State tags <S>, Material tags <M>, Action tags <A>, and Temporal tags <T>. These tags are embedded during prompt generation, ensuring that editing requirements have clear structural characteristics.

This tag specification offers the following advantages:

- **Automated Analysis:** Through clear tag boundaries, different types of editing operations can be programmatically extracted and analyzed. For example, we can quickly analyze the distribution of different object categories (via <O> tags) in the dataset, or assess the success rate of specific material transformations (via <M> tags).

Table 1
Comparison of benchmark dataset characteristics. The number of supported types indicates how many guidance types (text/sketch/example) the dataset supports. Patrick Star Bench is the only dataset that supports all three types of guidance.

Dataset	Source image	Text	Sketch	Example image	Mask	Groundtruth	Number of supported types
Text-guided							
EditBench [10]	✓	✓			✓	✓	1
TedBench [5]	✓	✓				✓	1
EditVal [13]	✓	✓				✓	1
IP2P [16]	✓	✓				✓	1
MagicBrush [15]	✓	✓			✓	✓	1
Emu Edit [9]	✓	✓					1
Example-guided							
COCOEE [11]	✓			✓	✓	✓	1
Sketch-guided							
SKETCH Dataset [12]			✓			✓	1
Patrick Star Bench	✓	✓	✓	✓	✓	✓	**3**

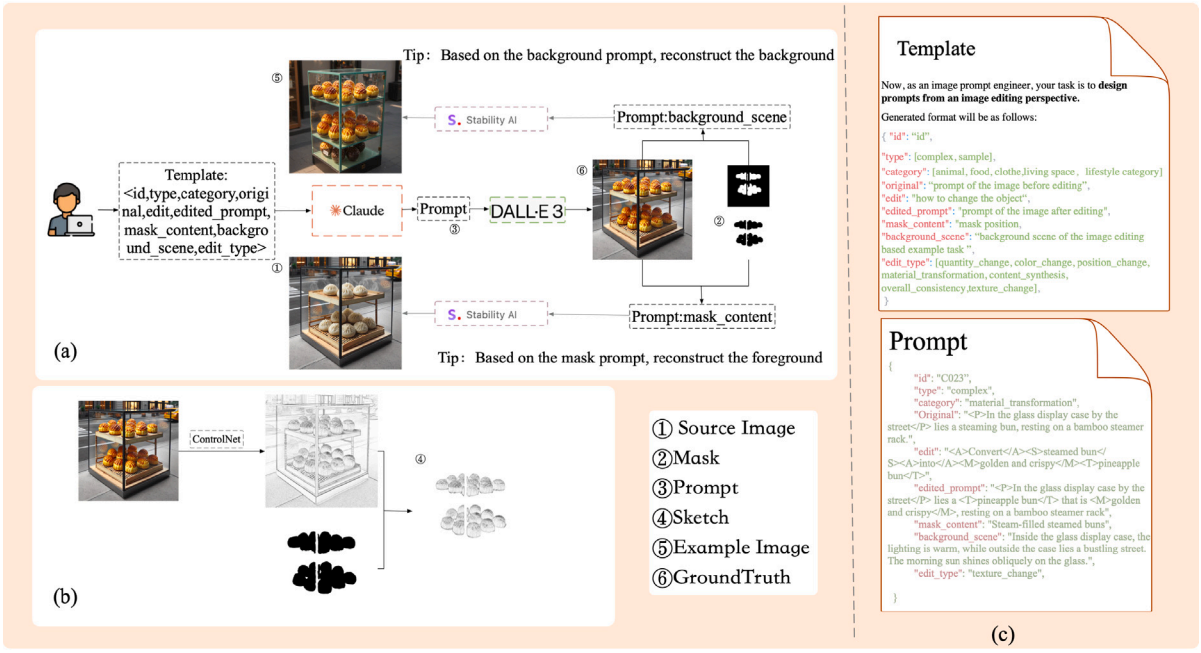

Fig. 4. Overview of Patrick Star dataset construction pipeline. (a) illustrates our multi-stage generation process: We first use Claude to generate structured prompts following our template format, then employ DALL-E 3 to create the ground truth image. After manual mask annotation or SAM [17] annotation on the ground truth, we utilize Stable Diffusion inpainting in two ways: combining the ground truth with mask and background_scene prompt to generate the source image (pre-editing state), and combining the ground truth with inverted mask and mask_content prompt to create the example image. (b) shows the sketch generation process, where we apply ControlNet's preprocessor to extract structural features from the source image and combine them with the mask to obtain the final line drawing. (c) presents our template structure for prompts, which includes comprehensive fields for image metadata, editing specifications, and contextual information. The complete dataset comprises seven essential components: source image, mask, text prompt, sketch, example image, ground truth, and mask content, collectively forming a comprehensive multi-modal benchmark for image editing evaluation.

- **Consistency Verification:** Using tag correspondence, we can automatically verify semantic consistency between pre- and post-editing descriptions. Particularly in complex editing tasks, these tags help track changes in key attributes.
- **Quality Control:** By enforcing tag specifications during the generation phase, we significantly reduce the need for manual review later in the process, improving the efficiency of dataset construction.

This tag specification serves as a crucial foundation for building Patrick Star Bench, providing reliable support for subsequent automated processing and analysis.

3.2. Multi-dimensional task taxonomy

To comprehensively evaluate image editing models' performance, we propose a two-level task classification system. Based on the complexity of editing operations and semantic levels, Patrick Star Bench categorizes tasks into Simple Tasks and Complex Tasks.

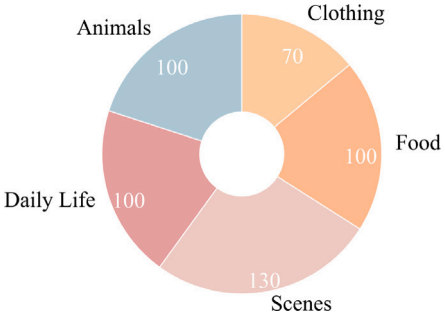

Fig. 5. Image category quantity chart.

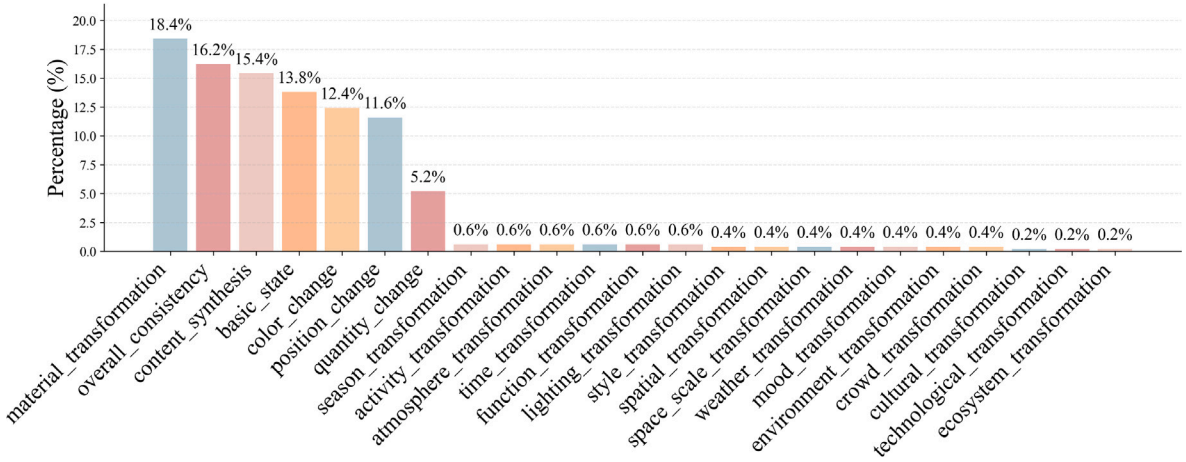

Fig. 6. Distribution of image editing operations in our dataset.

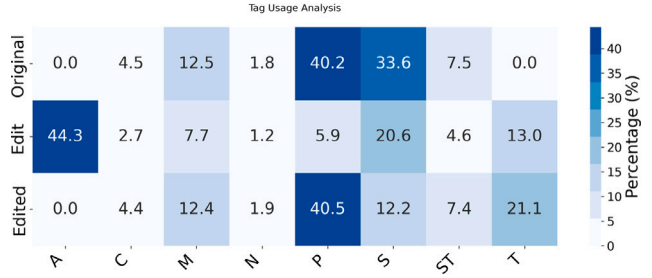

Fig. 7. A heatmap of the label and the changes in the quantities before, during, and after editing.

At the basic level, we define four fundamental editing operations: `quantity_change` focuses on precise modifications of object numbers in scenes; `color_change` addresses adjustments to basic appearance attributes; `position_change` evaluates models' capabilities in spatial layout modifications; and `basic_state` tests simple object attribute transformations. While these tasks are operationally simple, they require models to possess precise local editing capabilities.

Complex tasks examine models' ability to handle complex semantic transformations: `material_transformation` requires changing object physical properties while maintaining shape; `content_synthesis` tests models' ability to integrate multiple elements; `overall_consistency` evaluates scene coherence during large-scale editing; and `texture_ transformation` focuses on fine-grained surface feature modifications. These tasks not only demand accurate editing operations but also require maintaining natural contextual transitions. To be specific, the taxonomy categorizes the distribution of image editing operations, as shown in Fig. 6, while more illustrative examples of image editing are presented in Fig. 8.

3.3. Dataset construction pipeline

Our dataset construction approach combines both real-world photography and AI-generated content to ensure diversity and quality, as shown in Fig. 4.Specifically, our dataset consists of 100 high-quality images sourced from Unsplash, a copyright-free photography platform, and 400 AI-generated images. This hybrid approach leverages the authenticity of real photographs while maintaining scalability through generative models.

3.3.1. Dual-source data collection

For real-world images, we carefully selected 100 high-resolution photographs from Unsplash that serve as ground truth images. These images were processed through multimodal large language models to automatically generate initial descriptions following our template format, followed by manual refinement to ensure tag specification compliance. Masks for these images were generated using a hybrid approach: for regions identifiable by SAM, we automatically expanded the bounding boxes and applied masks accordingly, while intricate details were manually annotated by experts to ensure precise editing region definition. The corresponding source images and reference images were then created using Stable Diffusion inpainting models with varying prompts, maintaining consistency with our editing objectives.

For the AI-generated portion, we develop a systematic creation process starting with prompt design. The process begins with designing specific prompt templates for each task type that comply with our tag specification while capturing the core challenges of each editing operation. For example, in material transformation tasks, the template must explicitly specify the material characteristics before and after editing while maintaining other object attributes unchanged.

3.3.2. Content creation

The content creation phase employs different strategies based on the image source. For Claude AI-generated [18] content, we utilize the DALL-E3 [19] API with optimized parameters, requiring multiple generation attempts to obtain candidates that best align with the prompts. For Unsplash-sourced images, we employ Stable Diffusion inpainting to generate variations while preserving the high quality of the original photographs. In both cases, mask generation focuses on precise editing region definition, and sketch extraction utilizes ControlNet with optimized parameters.

3.3.3. Quality verification

Our quality verification process ensures consistency across both real and generated images. For Unsplash-sourced content, we pay particular attention to the quality of generated variations and their alignment with the original photographs. For AI-generated content, we focus on the consistency between different versions of the same scene. The automated system verifies tag consistency, cross-modal alignment, and editing region standardization, while task-specific verification ensures that each sample meets its unique requirements.

Through this hybrid approach, we created a dataset of 500 high-quality editing samples, combining the authenticity of real photographs with the scalability of AI generation. This combination provides a more comprehensive benchmark for evaluating image editing models, as it tests performance on both real-world photographs and AI-generated content. The dataset's diverse sources and standardized quality make it particularly valuable for assessing models' generalization capabilities across different image types and editing scenarios.

Fig. 8. Illustration of five categories of food image editing: material transformation, quantity change, overall consistency, basic state and color change. Each row demonstrates a specific editing type with its source image, binary mask, text prompt, generated sketch, example reference, and ground truth (GT) result.

Table 2
Patrick Star Bench's test results on different tasks.

Evaluation metric	Text-guided		Sketch-guided		Example-guided	
	Image inpainting-SD1.5	Image inpainting-SDXL [20]	Controlnet-SD2.1	Controlnet-SDXL [4]	Paint by example [11]	DesignEdit [21]
LPIPS↓	0.0978	**0.0678**	0.572	**0.473**	0.0948	**0.0284**
FID↓	22.270	**19.314**	45.156	**34.253**	22.785	**18.501**
Pre_error↓	0.126	**0.103**	–	–	0.126	**0.099**
CLIP_Score↑	65.9120	**71.2986**	–	–	**70.7843**	70.1917
SSIM ↑	0.8214	**0.8442**	0.3282	**0.4885**	0.8253	**0.8748**
Aesthetic Score ↑	**4.9033**	4.8567	**5.0463**	4.8899	4.9635	**5.1798**

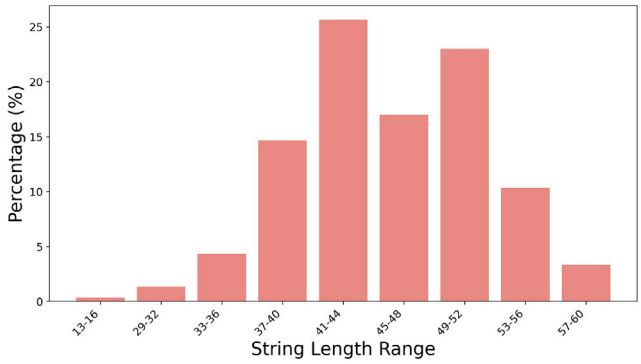

Fig. 9. Instruction string length.

3.4. Dataset statistics

To ensure comprehensive coverage and balance of the dataset, we conducted a detailed statistical analysis of various dataset features. As shown in Fig. 9, the instruction lengths exhibit diversity, with some being short for simple tasks like object replacement and color changes, while others are longer and more descriptive. These longer instructions typically require a certain level of detail to guide the model in making precise adjustments.

The source images in our dataset span across five main categories, as illustrated in Fig. 5: Natural Elements (25.0%), Daily Item Surfaces (23.0%), Clothing Parts (21.3%), Household Items (18.3%), and Food & Beverages (12.4%). This balanced distribution ensures the dataset's representativeness across different domains and editing scenarios. The relatively higher proportions of natural elements and surface textures reflect common editing requirements in real-world applications, while the inclusion of diverse categories enables comprehensive evaluation of models' generalization capabilities across different editing tasks.

4. Experiments

We conducted three groups of image editing experiments, totaling six tests covering both basic and optimized versions. All experiments were performed on an RTX 3090 GPU with 24 GB memory using identical hyperparameters, ensuring fair comparison conditions. For text-guided image editing, we employed SD1.5-Inpainting and SDXL-Inpainting models. Given the relatively limited work in sketch-guided image editing, we opted to use the ControlNet sketch generation models in both SD2.1 and SDXL versions. For example-guided image editing tasks, we compared Paint by Example with our proposed DesignEdit method. We evaluated the models using six evaluation metrics: LPIPS, FID and SSIM [22] for image quality assessment, Aesthetic Score [23] for image aesthetic evaluation, and CLIP_Score [24] and Pre_error for image content assessment.

The experimental results reveal significant performance variations across different approaches. As shown in Table 2, SDXL consistently outperforms SD1.5 and SD2.0 across most evaluation metrics, including LPIPS, FID, CLIP Score, and SSIM. This demonstrates that our dataset effectively differentiates the generation capabilities of different models. The ability to highlight these variations confirms the dataset's robustness in benchmarking image editing performance across multiple tasks and model architectures.

Human Evaluation We selected images generated by the six methods mentioned above and presented them to 100 participants. Participants were allowed to select multiple images that met the specified criteria. As shown in Fig. 10, they were asked to evaluate the images based on realism and alignment with the provided instructions. Inpainting-sd1.5, Inpainting-sdxl, Paint by Example, and DesignEdit received high scores for both realism and alignment. In contrast, ControlNet sdxl and ControlNet sd1.5 had lower scores, which demonstrates that our benchmark can effectively distinguish output images quality. Furthermore, the smaller differences in the ControlNet methods indicate that the benchmark can also capture subtle variations in performance. Overall, these results prove that the benchmark is effective and sensitive, providing valuable guidance for image editing tasks.

These comprehensive experimental results not only verify our dataset's applicability across different guidance modes but also demonstrate its effectiveness in evaluating and differentiating the performance of various methods. The consistent performance improvements observed in the optimized versions further confirm our dataset's discriminative capability and reliability, establishing a dependable benchmark for future model improvements and evaluations.

5. Conclusion and future work

This paper presents Patrick Star Bench, a comprehensive evaluation benchmark designed specifically for image editing tasks. Addressing the challenges of inconsistent evaluation standards and fragmented testing content in existing datasets, we have developed a systematic benchmark construction methodology grounded in scientific evaluation principles.

The key features and contributions of our benchmark dataset are significant. Notably, it is the first to provide unified support for three major editing paradigms: text-guided, sketch-guided, and example-guided editing. This integration establishes a reliable standard for evaluating model performance across cross-modal editing tasks. The dataset implements a rigorous quality control system that ensures prompt-image alignment, mask boundary precision, and sketch extraction clarity.

Through extensive validation experiments across six representative models, Patrick Star Bench has demonstrated excellent discriminative capability and reliability. The experimental results confirm that our benchmark can effectively assess performance differences between various editing methods while providing a solid foundation for future model improvements.

Looking ahead, we envision several promising directions for future research:

(1) Multi-turn Interactive Editing: Extending the benchmark to support evaluation of conversational image editing systems, where multiple rounds of user feedback and model responses are involved.

(2) Dynamic Assessment Metrics: Developing more sophisticated evaluation metrics that can capture the nuanced aspects of interactive editing processes.

(3) Temporal Consistency: Incorporating evaluation criteria for video editing tasks and sequential image modifications.

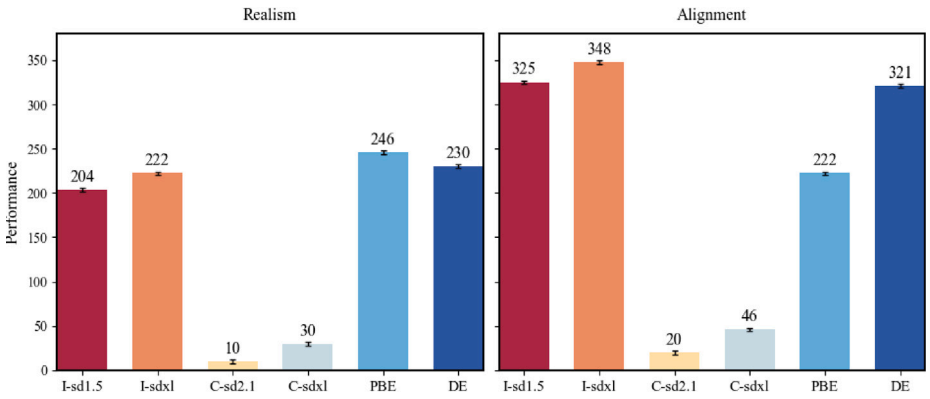

Fig. 10. Human evaluation of image realism and text–image alignment on Patrick Star.

These future developments aim to enhance the benchmark's functionality and broaden its applications in multimodal image editing research. We hope the foundation laid by Patrick Star Bench can contribute to the development of better evaluation methods for image editing technologies.

CRediT authorship contribution statement

Di Cheng: Writing – original draft, Conceptualization. **ZhengXin Yang:** Formal analysis, Conceptualization. **ChunJie Luo:** Methodology, Conceptualization. **Chen Zheng:** Data curation, Conceptualization. **YingJie Shi:** Writing – review & editing.

Declaration of competing interest

The author Chunjie Luo is the Assistant Editor in Chief and Chen Zheng is the Associate Editor for the journal BenchCouncil Transactions on Benchmarks, Standards and Evaluations and were not involved in the editorial review or the decision to publish this article. The other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to express our gratitude to the reviewers and editors for their constructive comments and suggestions. We also sincerely thank the teachers and students from the Institute of Computing Technology, Chinese Academy of Sciences, for their valuable feedback and support.

References

[1] Y. Huang, J. Huang, Y. Liu, M. Yan, J. Lv, J. Liu, W. Xiong, H. Zhang, S. Chen, L. Cao, Diffusion model-based image editing: A survey, 2024, arXiv:2402.17525.

[2] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[3] S. Basu, M. Saberi, S. Bhardwaj, A.M. Chegini, D. Massiceti, M. Sanjabi, S.X. Hu, S. Feizi, EditVal: Benchmarking diffusion based text-guided image editing methods, 2023, arXiv:2310.02426.

[4] P. Li, Q. Huang, Y. Ding, Z. Li, LayerDiffusion: Layered controlled image editing with diffusion models, 2023, arXiv:2305.18676.

[5] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, M. Irani, Imagic: Text-based real image editing with diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6007–6017.

[6] J. Zhuang, Y. Zeng, W. Liu, C. Yuan, K. Chen, A task is worth one word: Learning with task prompts for high-quality versatile image inpainting, in: European Conference on Computer Vision, Springer, 2025, pp. 195–211.

[7] Z. Geng, B. Yang, T. Hang, C. Li, S. Gu, T. Zhang, J. Bao, Z. Zhang, H. Li, H. Hu, et al., Instructdiffusion: A generalist modeling interface for vision tasks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 12709–12720.

[8] S. Yang, L. Zhang, L. Ma, Y. Liu, J. Fu, Y. He, Magicremover: Tuning-free text-guided image inpainting with diffusion models, 2023, arXiv preprint arXiv:2310.02848.

[9] S. Sheynin, A. Polyak, U. Singer, Y. Kirstain, A. Zohar, O. Ashual, D. Parikh, Y. Taigman, Emu edit: Precise image editing via recognition and generation tasks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8871–8879.

[10] S. Wang, C. Saharia, C. Montgomery, J. Pont-Tuset, S. Noy, S. Pellegrini, Y. Onoe, S. Laszlo, D.J. Fleet, R. 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, 2023, pp. 18359–18369.

[11] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, F. Wen, Paint by example: Exemplar-based image editing with diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18381–18391.

[12] M. Eitz, J. Hays, M. Alexa, How do humans sketch objects? ACM Trans. Graph. 31 (4) (2012) 1–10.

[13] S. Basu, M. Saberi, S. Bhardwaj, A.M. Chegini, D. Massiceti, M. Sanjabi, S.X. Hu, S. Feizi, Editval: Benchmarking diffusion based text-guided image editing methods, 2023, arXiv preprint arXiv:2310.02426.

[14] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, P. Dollár, Microsoft COCO: Common objects in context, 2015, arXiv:1405.0312.

[15] K. Zhang, L. Mo, W. Chen, H. Sun, Y. Su, Magicbrush: A manually annotated dataset for instruction-guided image editing, Adv. Neural Inf. Process. Syst. 36 (2024).

[16] Q. Guo, T. Lin, Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 6986–6996.

[17] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment anything, 2023, arXiv:2304.02643.

[18] Anthropic, Claude AI, 2025, https://claude.ai. (Accessed 29 January 2025).

[19] L. Zhang, A. Rao, M. Agrawala, Adding conditional control to text-to-image diffusion models, in: IEEE International Conference on Computer Vision, ICCV.

[20] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, S. Ermon, SDEdit: Guided image synthesis and editing with stochastic differential equations, 2022, arXiv:2108.01073.

[21] Y. Jia, Y. Yuan, A. Cheng, C. Wang, J. Li, H. Jia, S. Zhang, DesignEdit: Multi-layered latent decomposition and fusion for unified & accurate image editing, 2024, arXiv preprint arXiv:2403.14487.

[22] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.

[23] C. Schuhmann, R. Beaumont, LAION-AI aesthetic-predictor, 2022, URL https://github.com/LAION-AI/aesthetic-predictor. (Accessed 27 March 2025).

[24] J. Hessel, A. Holtzman, M. Forbes, R.L. Bras, Y. Choi, CLIPScore: A reference-free evaluation metric for image captioning, 2022, arXiv:2104.08718.