Redesign Your Space with Image Inpainting Model

2024 CVPDL Final Project Group 27

B10502010 Wei-Chin Wang*, B10505029 Ray-Ting Huang*, B10705005 Szu-Ju Chen†,
R12522508 Pei-Lin Hou‡, R12945070 Kao-Tiem Liu§,
*Department of Electrical Engineering, NTU
‡Department of Mechanical Engineering, NTU
§Department of Biomedical Electronics and Bioinformatics, NTU
†Department of Computer Science and Information Engineering, NTU

***Abstract*—This project presents an innovative tool for interior design that enables users to effortlessly replace objects in images and optimize scenes. By integrating state-of-the-art techniques, such as SegVG for segmentation and PowerPaint for image inpainting, the tool utilizes text-based prompts to guide object replacement while ensuring semantic consistency and visual coherence. We fine-tune the segmentation model on a custom indoor dataset and evaluate the overall performance using CLIP score metrics. Additionally, we conduct various experiments and compare the proposed approach with other potential methods. Our results demonstrate that the proposed pipeline outperforms existing techniques, including BrushNet. Despite challenges with mask quality and incomplete training data, we propose solutions and outline future work, such as expanding the dataset and exploring self-supervised learning to further enhance performance.**

***Keywords*—Visual Grounding, Obeject Detection, Transformers, Object Removal, Image Inpainting, Diffusion Model**

I. Introduction

The increasing popularity of text-to-image (T2I) models has led to significant advancements in visual grounding for understanding user queries and image inpainting for enhancing image quality based on user requests. These technologies deliver efficient and highly effective results. Through our exploration of various application fields, we identified a notable challenge faced by interior designers: the reliance on manual editing and the lack of operational flexibility to accommodate customer preferences and rapidly changing design trends. To address this issue, our project introduces an efficient and practical design tool aimed at assisting both designers and customers in transforming their creative inspirations into tangible design outcomes. This tool emphasizes a user-friendly interface, intuitive prompts, and short processing times.

In this project, we integrate state-of-the-art (SOTA) models, including SegVG [4] and PowerPaint [8], to develop a comprehensive and efficient workflow. The process begins with bounding-box segmentation based on natural language inputs using a fine-tuned SegVG model and a nearest-exact interpolation algorithm. The generated mask is then processed by the inpainting model, PowerPaint, which replaces the target object with the desired object while preserving the surrounding background. Our workflow eliminates the need for manual selection of masked areas, allowing users to achieve results effortlessly by providing just one image and two prompts. Our contributions are summarized as follows:

- We propose a tool that integrates SOTA models to perform design operations effortlessly and effectively.
- We explore and implement various optimization techniques for mask generation to improve segmentation accuracy and practical usability.
- We conduct extensive comparisons and evaluations of PowerPaint and BrushNet to ensure the workflow delivers natural and consistent results.
- Our workflow offers a novel solution tailored for interior design and related fields, addressing current challenges and serving as a reference for future applications.

II. Related Work

Visual Grounding. Visual grounding focuses on localizing a target object within an image based on a free-form natural language description. This capability is crucial for understanding specific user requests, especially when multiple candidates exist within an image. Recent approaches employ one-stage, two-stage, and transformer-based methods to generate object masks in various forms, such as bounding boxes or segmentation masks. We experimented with both techniques, leveraging two recent advancements:

- **SegVG.** SegVG [4] introduces an innovative approach by utilizing pixel-level details within box annotations as segmentation signals. It employs a Multi-layer Multi-task Encoder-Decoder architecture and Triple Alignment strategy to iteratively leverage annotations for both box-level regression and pixel-level segmentation. Additionally, SegVG reduces domain discrepancies among the query, text, and vision inputs, enhancing segmentation accuracy and adaptability.
- **EVF-SAM.** EVF-SAM [5] builds upon SAM [6] by addressing language understanding and text-prompted queries. It integrates a Multimodal Encoder with Early Vision-Language Fusion to convert multimodal inputs into prompt embeddings for SAM. This outperforms traditional text-encoders and large language models, providing a more stable and efficient training paradigm for text-prompted SAM applications. capabilities

Image Inpainting. Image inpainting involves restoring or reconstructing missing or corrupted parts of an image in a visually coherent and plausible manner. It finds applications in

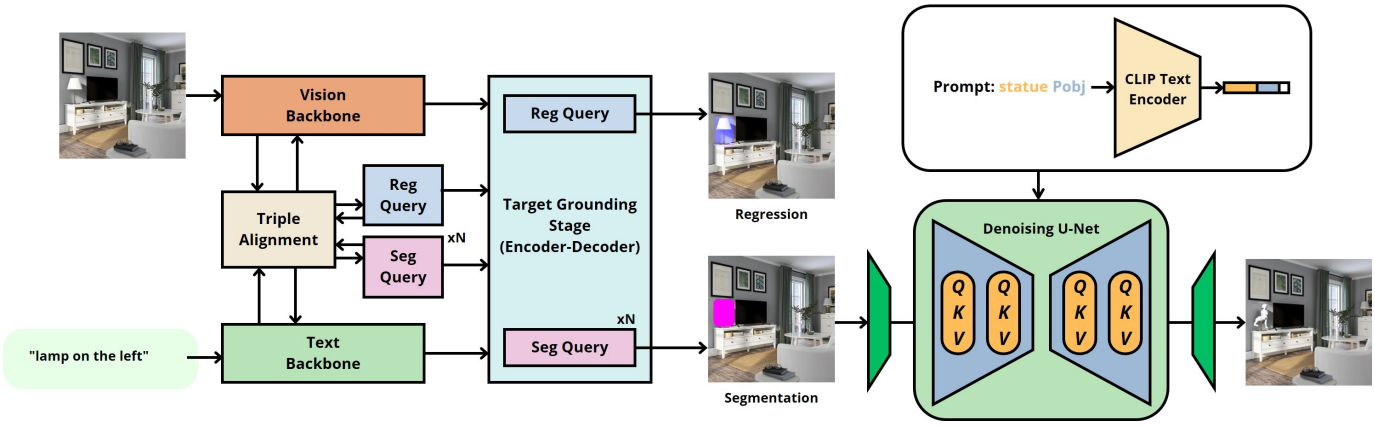

Fig. 1. workflow of our integrated tool

photo restoration, object removal, and content-aware editing. Modern techniques leverage advanced generative models, such as diffusion models, which use contextual, structural, and textual features to ensure semantic and structural coherence. Our work builds upon two recent contributions:

- **PowerPaint.** PowerPaint [8] addresses the challenges of context-aware filling and object synthesis by introducing learnable task prompts tailored for distinct objectives. It supports tasks like text-guided object inpainting, context-aware image inpainting, and shape-guided object inpainting. By fine-tuning a pre-trained diffusion model with task-specific prompts and techniques such as prompt interpolation and classifier-free guidance, PowerPaint achieves high-quality results. It excels in aligning generated content with user-provided masks and textual descriptions, ensuring contextual coherence and adaptability across various inpainting tasks.
- **BrushNet.** BrushNet [7] employs a dual-branch architecture that separates masked image feature processing from noisy latent generation. This design improves context-aware feature extraction and pixel-level control. By integrating hierarchical feature extraction and removing text cross-attention in the masked image branch, BrushNet resolves challenges like semantic inconsistencies and low image quality in complex masked regions. As a result, it delivers state-of-the-art performance with enhanced semantic alignment and consistency.

III. Methodology

The workflow of our integrated tool is illustrated in Fig. 1 and consists of four main stages: (1) **User Input**: The user provides an input image and two text prompts—one specifying the object to be removed and the other describing the replacement object. (2) **Segmentation Model**: The model processes the input image and the ”remove object” prompt to generate a precise, pixel-level mask for the specified object. (3) **Inpainting Model**: The model utilizes the generated mask and the ”replacement object” prompt to produce the final result. (4)**Result Delivery**: The final edited image is returned to the user, providing a visually refined output that reflects their specifications.

A. Target Object Mask on Text-based Prompt

To ensure accurate object masking from user-provided text prompts, we utilize segmentation models capable of referring object understanding. Among the evaluated approaches, SegVG [4] outperforms EVF-SAM [5] in providing higher coverage of the target object. Specifically, SegVG generates segmentation signals in a bounding-box shape that encompass the entire object, whereas EVF-SAM's segmentation occasionally omits parts of the object. For instance, as illustrated in Fig. 2, SegVG successfully masks the entire chair, while EVF-SAM leaves the left leg unmasked. Since the subsequent inpainting process is highly sensitive to unmasked regions, we chose SegVG as our base model to minimize the risk of errors.

In the segmentation process, SegVG processes the ”remove object” prompt through its pre-trained transformer encoder, DETR, while the input image is processed via the ResNet vision backbone. These components generate semantic features from the text prompt and visual features from the image, which are fused in the Triple Alignment module. This produces two query vectors: (1) Reg Query: representing the predicted bounding box. (2) Seg Query: representing the pixel-level segmentation mask. Using these two queries and the multi-layer multi-task encoder-decoder architecture, the model learns to localize the target object by aligning text semantics with image features, ensuring precise and comprehensive segmentation, forming a robust foundation for subsequent inpainting.

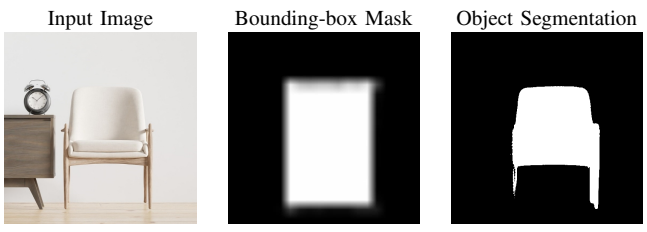

Fig. 2. Segmentation Mask Comparison

B. Image Inpainting on User Request

Using the segmentation mask and the user-provided target object prompt, the inpainting model replaces the specified object. PowerPaint [8] enhances inpainting quality by introducing learnable task prompts, P_{obj}, tailored for specific objectives among other models. These prompts are trained using object bounding boxes as masks, which are appended to the text descriptions of the masked regions. This technique enables the model to synthesize objects by integrating textual cues with precise spatial constraints, optimizing task-specific prompts to align with the input conditions.

In this stage, the target object prompt is processed by the CLIP Text Encoder, with P_{obj} concatenated to improve the model's semantic understanding. The enriched textual semantic features, along with the segmentation mask from the previous stage, are inputted into the Denoising U-Net. The U-Net's multi-layer attention mechanism reconstructs and denoises the specified region, generating a new object that aligns with the user's description. Simultaneously, it preserves the background information, ensuring the output remains semantically consistent and visually coherent.

IV. Experiments

A. Indoor Dataset

To fine-tune our visual grounding segmentation model for indoor scenarios, we curated a dataset by combining 1500 images from the OpenImages Dataset v7 [1] and 5000 images from RefCOCO [2]. These datasets were filtered to include images depicting indoor environments and objects. Additionally, we utilized InstructDET [3] to generate referring annotations for the OpenImages [1] data, enhancing the dataset with high-quality, context-aware labels.

InstructDET, a data-centric method for Referring Object Detection (ROD) that employs Vision-Language Models (VLMs) and Large Language Models (LLMs) to generate natural language instructions for target object localization. It creates two types of prompts: (1) Global prompt: diverse expressions describing the properties, categories, and relationships of a single object. (2) Local Prompt: informative expressions that are closely tied to the specific object of interest. To ensure robustness, InstructDET uses a CLIP model as a verifier, calculating global and local prompt scores. This mitigates CLIP's tendency to favor large objects and ensures the generated prompts are accurate and contextually relevant.

B. Segmentation Fine-tune

To achieve domain-specific results for interior design tasks, we fine-tuned the SegVG [4] model using the curated indoor dataset. We selected ResNet101, pre-trained on ReferItGame [2], as the vision backbone and initialized the SegVG model with published checkpoints fine-tuned on ReferItGame.

We experimented with four fine-tuning configurations, using a training-validation split of 90:10:

- RefCOCO: includes people and indoor objects
- RefCOCO: focuses only indoor objects
- InstructDET: generated from the filtered OpenImages
- Combined Dataset: RefCOCO Indoor Objects and InstructDET annotations

As shown in Fig. 3, the training loss converged after 40 epochs and the validation accuracy stabilized at the same point. Also as the Table I illustrates that the finetuning task on RefCOCO data have the best accurate mask performance on the validation data. Bases on these results, we selected the fine-tuned RefCOCO model as the primary took for mask generation.

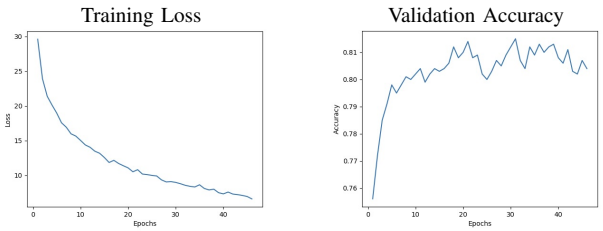

Fig. 3. Fine-tuned training loss and validation accuracy

Fine-tuned Data	Improve training loss	Improve accuracy	Best accuracy
RefCOCO	-23.04	+0.06	**0.81**
RefCOCO (Indoor only)	-27.72	+0.077	0.75
InstructDet	**-56.44**	**+0.21**	0.61
Combined	-35.78	0.15	0.60

TABLE I: Model Training Results

C. Segmentation Mask

SegVG [4] generates a flattened vector of size 400 to represent the masked area, which must be resized to match the input image dimensions for the subsequent inpainting process. We evaluated four interpolation algorithms for resizing, as detailed below:

- Nearest: Assigns the value of the nearest neighbor.
- Nearest-exact: Refines the nearest neighbor approach for better alignment.
- Bilinear: Performs linear interpolation in two dimensions.
- Bicubic: Uses cubic interpolation in two dimensions for smoother transitions.

Fig. 4 illustrates the masks produced by each algorithm. Through human evaluation and experimental trials, we found that bilinear interpolation provided the best masking results. It produced smoother edges with beneficial blurring effects, aiding the inpainting process. Additionally, bilinear interpolation yielded a more pronounced contrast between masked and unmasked regions compared to bicubic interpolation.

D. Metric

The CLIP Score [9] is a similarity-based metric used to assess semantic consistency between a generated image

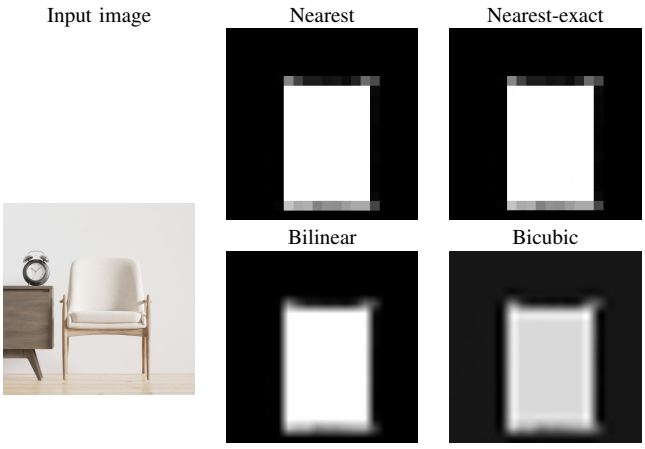

Fig. 4. Different interpolation mask

and its corresponding textual description. Renowned for its performance in multimodal alignment tasks, the CLIP Score provides a robust quantitative evaluation for our project.

For evaluation, we employ Google Gemini to generate descriptions of the original input image. These descriptions are manually adapted to reflect the user-specified object modifications, resulting in a final description for the generated image. This dual approach ensures alignment with the user's input.

A higher CLIP Score indicate greater semantic alignment, ensuring that the generated object and scene are consistent with user requests.

E. Inpainting Model Comparison

Given the rapid advancements in inpainting models, we conducted a comparative study with BrushNet [7], a state-of-the-art inpainting model that employs a dual-branch architecture to minimize noise during the generation process. To ensure a fair comparison, we followed the same workflow and use identical masks generated by the fine-tuned SegVG [4] model.

The comparison was based on the CLIP score [9], which evaluates semantic consistency, visual coherence, and inpainting quality. As shown in Fig. 5, Fig. 6, Fig. 7, the PowerPaint [8] model demonstrated competitive performance and consistently outperformed BrushNet. While BrushNet excels in pixel-level precision, it occasionally struggles to maintain global coherence in complex scenes. In contrast, PowerPaint integrates the generated object more smoothly into the surrounding environment, resulting in superior semantic alignment. These findings underscore the advantage of incorporating P_{obj} into the semantic feature representation in PowerPaint, enabling it to preserve generated quality effectively. Moreover, this validates our workflow's capability to address complex indoor design challenges, establishing PowerPaint as a robust solution for such tasks.

F. Qualitative Results

We present a comprehensive set of results to evaluate the performance of PowerPaint [8] and BrushNet [7] across varying levels of self-defined mask quality. The results are grouped into three categories: (1) accurate and precise masks, (2) incomplete masks that fail to fully cover the intended bounding box, and (3) incorrect masks with significant misalignment or errors. The results are shown in Fig. 5, Fig. 6, Fig. 7, and the overall quantitative CLIP score is illustrated at Table II.

Mask Quality	PowerPaint CLIP score	BrushNet CLIP score
correct mask	**14.84**	10.36
incomplete mask	**17.70**	10.37
incorrect mask	14.75	**17.98**
All	**16.23**	12.44

TABLE II: Model Comparison on CLIP Score

V. Discussions

Despite integrating state-of-the-art models and implementing various enhancements, certain challenges still affect the overall performance of our tool. Specifically, the inpainting process heavily depends on the quality of the segmentation mask generated in the preceding stage. Through our analysis of undesired outcomes, we identified two primary issues in the segmentation task:

A. Mask quality issue

The accuracy of the inpainting process is highly sensitive to incomplete or inaccurate masks. In cases where the mask fails to fully cover the target object, the inpainting model tends to preserve the unmasked region, resulting in minimal modification of the original image. For example, as shown in Fig. 6, especially the first two columns, incomplete masking of the upper part of the lamp and the body part of the fan leads to generated images where remnants of the original objects are still visible.

To address this issue, we propose the following hypotheses: (1) Increasing Mask Resolution: Utilizing a larger mask vector may reduce errors by providing finer details during interpolation. (2) Thresholding Mask Values: Applying a threshold to generate binary masks (values of 0 or 255) could enhance the inpainting model's confidence in the mask, leading to improved object replacement

B. Limitation on training dataset

Another limitation arises from incorrect object replacements or inaccurate localization of target regions, which we attribute to gaps in the segmentation model's ability to interpret certain prompts and referring conditions. For instance, as illustrated in Fig. 7, the model fails to correctly identify the "printer machine," resulting in an entirely misplaced masked region on the poster. Similarly, it struggles to accurately locate the "left pillow," leading to misaligned object generation that affects a larger portion of the bed.

We believe these issues can be mitigated by augmenting the training dataset with more diverse and precise indoor scenes. A richer dataset would improve the segmentation model's

ability to interpret complex prompts and ensure accurate localization of target regions, ultimately enhancing the inpainting outcomes.

VI. Conclusion and Future Work

This project introduces an efficient design tool that empowers users to transform their creative inspirations into tangible interior design outcomes with ease. By integrating advanced technologies such as SegVG [4] and PowerPaint [8], we have developed a comprehensive workflow that enables seamless object replacement and scene optimization through intuitive, text-based prompts. The tool leverages natural language understanding and state-of-the-art inpainting techniques to produce visually refined, semantically consistent results that blend smoothly with the existing background. Despite these advancements, our findings highlight areas for improvement. A primary focus moving forward is to enhance the quality of referral mask generation by exploring novel methodologies, refining existing algorithms, and expanding the dataset to include a broader range of diverse interior objects. Additionally, we aim to investigate the use of self-supervised learning and Generative Adversarial Networks (GANs) to further boost the model's learning capacity and adaptability, particularly in scenarios with limited data availability. Finally, we plan to extend this approach to other design domains, such as outdoor landscapes or virtual scene modeling, to evaluate its versatility and practical value across diverse applications.

VII. Contribution Attribution

A. Our original approach

- Definition of the project's data domain
- Creation of the referral-annotated dataset on OpenImages
- Fine-tuning task on the segmentation model
- Development of the user interface for our tool

B. External Source

- RefCOCO dataset [2]
- SegVG model code
  https://github.com/WeitaiKang/SegVG
- PowerPaint model code
  https://github.com/open-mmlab/PowerPaint
- BrushNet model code
  https://github.com/TencentARC/BrushNet

VIII. Team Contribution

B10502010 Wei-Chin Wang: Indoor dataset filtering and generation, CLIP score evaluation, presentation

B10505029 Ray-Ting Huang: Image inpainting model study and implementation, UI interface development

B10705005 Szu-Ju Chen: Referral object segmentation study and implementation, model fine-tuning, report writing

R12522508 Pei-Lin Hou: Evaluation metric study, slides organization, presentation

R12945070 Kao-Tiem Liu:

IX. Demo Video

https://youtu.be/mH3705oWZzA

References

[1] Krasin I., Duerig T., Alldrin N., Ferrari V., Abu-El-Haija S., Kuznetsova A., Rom H., Uijlings J., Popov S., Kamali S., Malloci M., Pont-Tuset J., Veit A., Belongie S., Gomes V., Gupta A., Sun C., Chechik G., Cai D., Feng Z., Narayanan D., Murphy K. OpenImages: A public dataset for large-scale multi-label and multi-class image classification, 2017.

[2] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In Proceed- ings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 787– 798. 2014.

[3] Ronghao Dang, Jiangyan Feng, Haodong Zhang, Chongjian Ge, Lin Song, Lijun Gong, Chengju Liu, Qijun Chen, Feng Zhu, Rui Zhao, et al.Instructdet: Diversifying referring object detec- tion with generalized instructions. arXiv preprint arXiv:2310.05136. 2023.

[4] Kang, W., Liu, G., Shah, M., Yan, Y. "SegVG: Transferring Object Bounding Box to Segmentation for Visual Grounding." In European Conference on Computer Vision. Springer, Cham, 2025.

[5] Zhang, Y., Cheng, T., Hu, R., Liu, L., Liu, H., Ran, L., Chen, X., Liu, W., Wang, X. (2024). "EVF-SAM: Early Vision-Language Fusion for Text-Prompted Segment Anything Model." arXiv preprint arXiv:2406.20076.

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

[7] Ju, X., Liu, X., Wang, X., Bian, Y., Shan, Y., Xu, Q. (2024). Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. arXiv preprint arXiv:2403.06976.

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

[9] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I. (2021). "Learning Transferable Visual Models From Natural Language Supervision." In International conference on machine learning (pp. 8748-8763). PMLR.

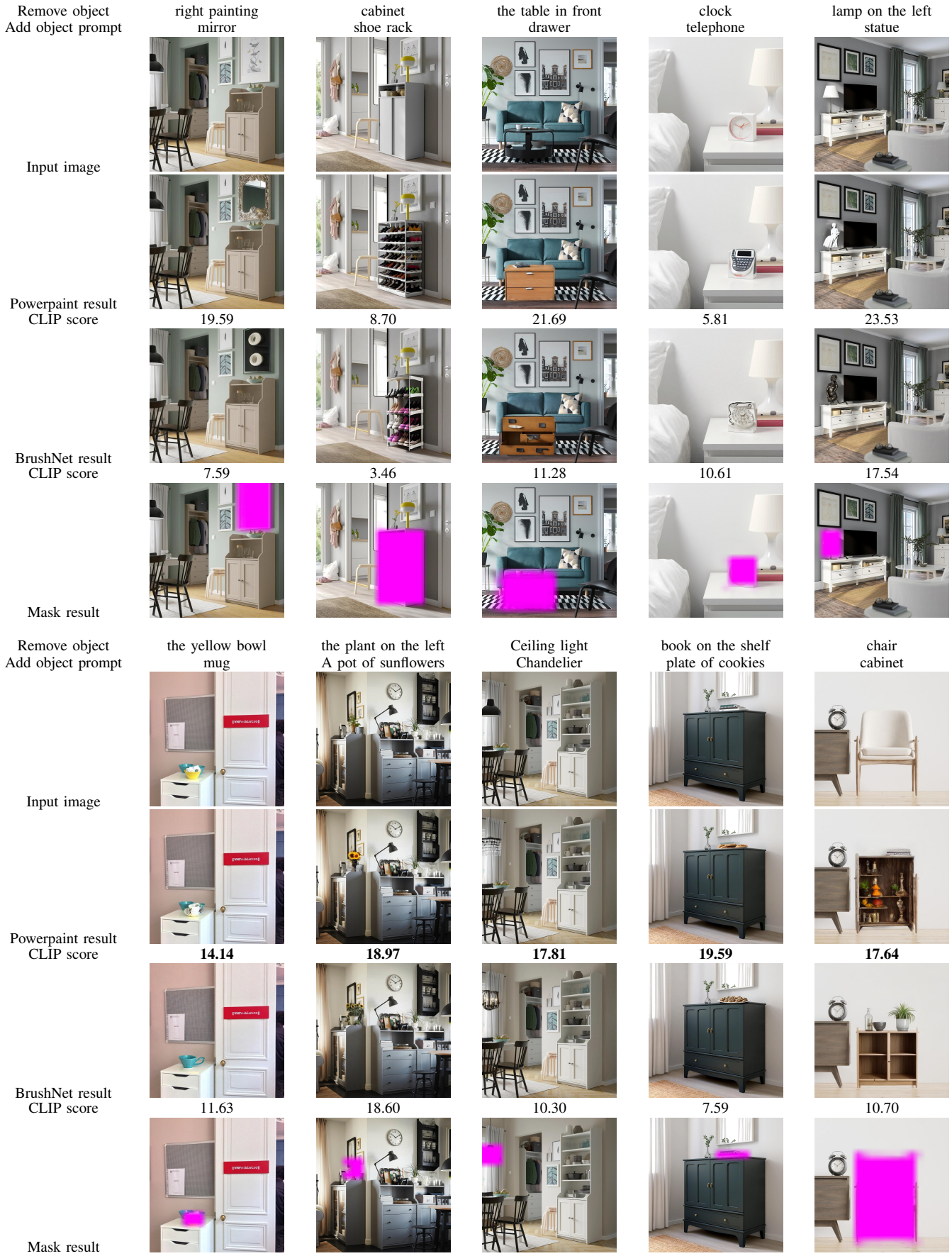

Fig. 5. Qualitative Result of Correct Mask Generation

Fig. 6. Qualitative Result of Incomplete Mask Generation

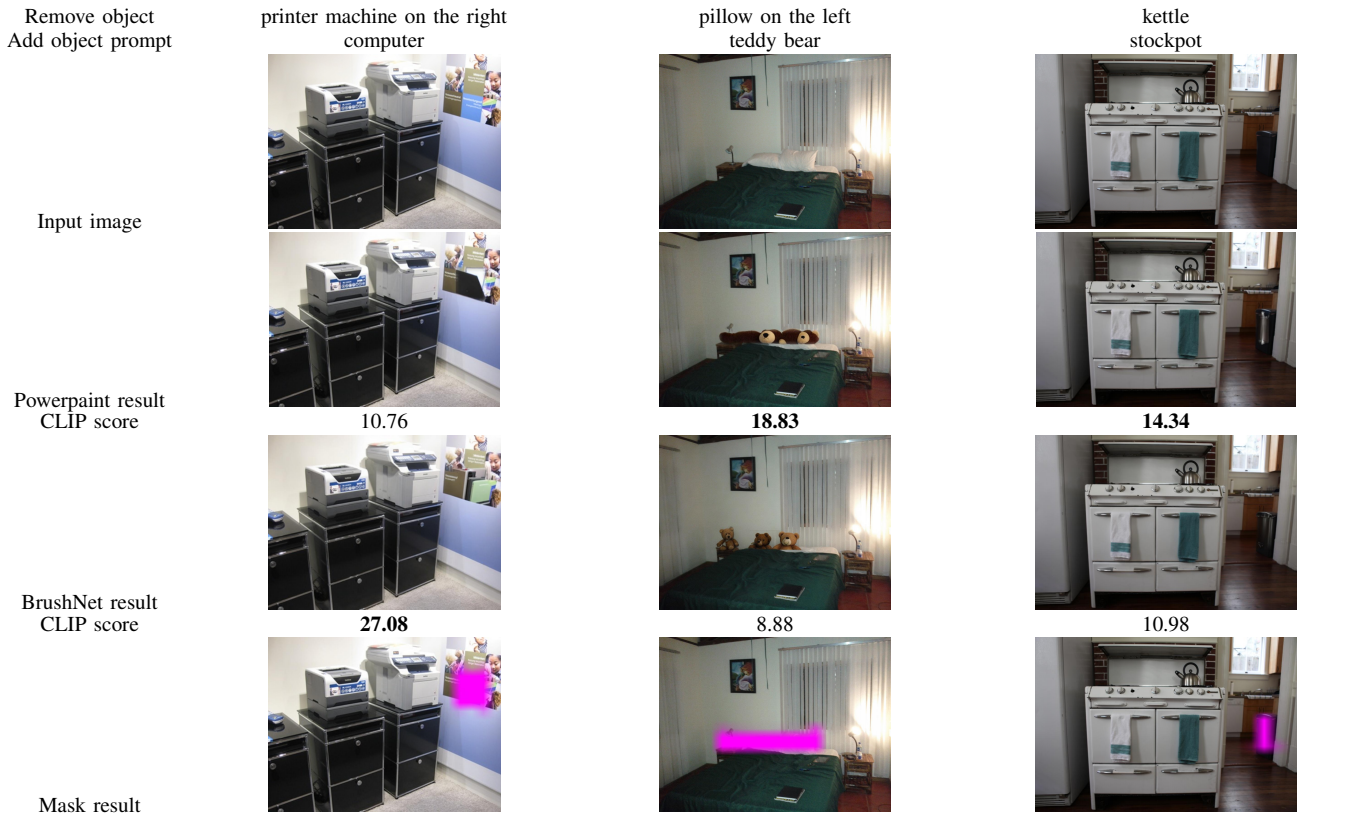

Fig. 7. Qualitative Result of Incorrect Mask Generation