ANYWHERE: A MULTI-AGENT FRAMEWORK FOR RELIABLE AND DIVERSE FOREGROUND-CONDITIONED IMAGE INPAINTING

Tianyidan Xie
State Key Laboratory of Novel Software Technology
Nanjing University
sealical@outlook.com

Rui Ma
Jilin University

Qian Wang[*]
China Mobile Communications Group Co.,Ltd

Xiaoqian Ye
China Mobile Communications Group Co.,Ltd

Feixuan Liu
Larkagent AI

Ying Tai
State Key Laboratory of Novel Software Technology
Nanjing University

Zhenyu Zhang
State Key Laboratory of Novel Software Technology
Nanjing University

Zili Yi[*]
State Key Laboratory of Novel Software Technology
Nanjing University

April 30, 2024

ABSTRACT

Recent advancements in image inpainting, particularly through diffusion modeling, have yielded promising outcomes. However, when tested in scenarios involving the completion of images based on the foreground objects, current methods that aim to inpaint an image in an end-to-end manner encounter challenges such as "over-imagination", inconsistency between foreground and background, and limited diversity. In response, we introduce Anywhere, a pioneering multi-agent framework designed to address these issues. Anywhere utilizes a sophisticated pipeline framework comprising various agents such as Visual Language Model (VLM), Large Language Model (LLM), and image generation models. This framework consists of three principal components: the prompt generation module, the image generation module, and the outcome analyzer. The prompt generation module conducts a semantic analysis of the input foreground image, leveraging VLM to predict relevant language descriptions and LLM to recommend optimal language prompts. In the image generation module, we employ a text-guided canny-to-image generation model to create a template image based on the edge map of the foreground image and language prompts, and an image refiner to produce the outcome by blending the input foreground and the template image. The outcome analyzer employs VLM to evaluate image content rationality, aesthetic score, and foreground-background relevance, triggering prompt and image regeneration as needed. Extensive experiments demonstrate that our Anywhere framework excels in foreground-conditioned image inpainting, mitigating "over-imagination", resolving foreground-background discrepancies, and enhancing diversity. It successfully elevates foreground-conditioned image inpainting to produce more reliable and diverse results. See our project page at https://anywheremultiagent.github.io.

---

[*]Corresponding author

Figure 1: Left: Three issues faced by existing approaches for foreground-conditioned image inpainting: limited diversity, "over-imagination", and foreground-background inconsistency. Please note that the red circles highlight the regions with "over-imagination". Right: the outcomes of our multi-agent framework.

1 Introduction

The rapid advancement of diffusion models has revolutionized image inpainting [1]. Text-to-image generation models enable users to control the diffusion process with textual or multi-modal information [2, 3, 4, 5], thus allowing for more personalized image inpainting by incorporating text or other modalities as additional cues [6, 7, 2]. Meanwhile, researchers are tackling more challenging inpainting tasks such as background-conditioned object hallucination or foreground-conditioned image completion [6, 7, 8, 9, 10, 11, 12, 13]. Specifically, HD-painter [9] introduces a training-free method that precisely adheres to prompts and seamlessly scales to high-resolution image inpainting, by introducing a novel Prompt-Aware Introverted Attention (PAIntA) layer. BrushNet [11] presents a novel plug-and-play dual-branch model designed to integrate pixel-level masked image features into any pre-trained diffusion model, ensuring coherent and improved image inpainting results. LayerDiffussion [12] facilitates large-scale pre-trained latent diffusion models to generate single transparent images or multiple transparent layers by learning a "latent transparency", which enables foreground- or background-conditioned image inpainting.

However, with regards to foreground-conditioned image inpainting, existing methods still encounter issues such as "over-imagination", foreground-background inconsistency, and limited diversity: see Figure 1. "Over-imagination" refers to the generation of redundant or excessive contents around the foreground object, compromising foreground integrity (e.g., adding unnecessary area to a chair). Secondly, foreground-background inconsistency involves placing the foreground object in an inappropriate or irrelevant environment (e.g., slippers on a campfire), inconsistent viewpoint or spatial relations (e.g., a bird's-eye-view object in a horizontal-view background, a floating watch in a room), and inappropriate relative size of the foreground object and background setting (e.g., a cup is bigger than the desk). Thirdly, limited diversity refers to the inability of the inpainting model to generate diverse results, resulting in predominantly uniform or visually similar backgrounds.

When tackling these challenges, we note that end-to-end models often struggle to comprehend foreground contents accurately, lack the capability to fill in missing information creatively, and lack mechanisms to prevent "over-imagination". To address these challenges, we introduce Anywhere, a novel multi-agent framework that employs a sophisticated pipeline comprising various agents such as VLM [14, 15, 16, 17], LLM [18, 16], SDXL [6, 7], and ControlNet [2]. This framework consists of three components: the prompt generation module, the image generation module, and the outcome analyzer. The prompt generation module conducts semantic analysis of the input foreground image, utilizing VLM to predict relevant language descriptions and LLM to recommend optimal language prompts. These prompts are further used to guide the image generation module, ensuring the avoidance of irrelevant content generation and promoting diversity. In the image generation module, we utilize a text-guided canny-to-image generation model, such as the ControlNet Canny model [2], to create a template image based on the edge map of the foreground image and language

prompts. Additionally, we employ a copy-and-paste tool to preserve foreground integrity and image blending agents to ensure foreground-background harmony. Moreover, a text-guided image inpainting model is used as a re-inpainting agent to address instances of "over-imagination", when "over-imagination" is detected by the auto-detection tool. The outcome analyzer utilizes VLM to assess image content rationality, aesthetic score, and foreground-background relevance, triggering prompt and image regeneration as needed. The outcome analyzer can be used through multiple rounds of iteration, ensuring more reliable results with the feedback mechanism.

Extensive experiments demonstrate the effectiveness of our Anywhere framework in foreground-conditioned image inpainting, mitigating "over-imagination" and foreground-background discrepancies, and enhancing diversity. Qualitative and quantitative evaluations demonstrate that our multi-agent framework are significantly more reliable and diverse than existing end-to-end image inpainting approaches.

To sum up, the major contributions of this paper include:

- We introduce a novel multi-agent framework that incorporates advanced VLM, LLM, and image generation models to address the task of foreground-conditioned image inpainting, significantly surpassing existing end-to-end approaches in generating reliable and diverse inpainting results.

- We introduce a novel mechanism for the auto-detection of "over-imagination" and image template re-inpainting to mitigate the "over-imagination" issue.

- We employ a novel multi-round iterable outcome analyzer to trigger regeneration of the language prompts and inpainting results for more reliable outputs.

2 Related work

2.1 Diffusion-based Controllable Image Generation

Stable diffusion, a prominent open-source text-to-image (T2I) model, has witnessed rapid advancements recently. However, user requirements often extend beyond textual descriptions. Researchers have attempted to add additional control signals to influence the diffusion process, such as adding subject images [19, 20, 21, 22, 23, 24] and style [25, 26]. Some studies focus on extra specific control signals such as layout condition [4, 5], edge map [2, 3], segmentation mask [27, 2], viewpoint [28, 29, 30]. LayerDiffusion [12] is concerned with generating images on transparent layers, and the resulting foreground or background can be used as a control condition to guide the text-to-image diffusion process.

2.2 Diffusion-based Image Inpainting

Image inpainting is a pivotal task in computer vision, focusing on the restoration of masked regions based on surrounding unmasked content. Recent advancements in diffusion modeling have significantly propelled the field of inpainting forward. Notable techniques include Palette [31] and Repaint [32], which leverage the original image alongside the unmasked regions to enhance denoising. Blended Diffusion [33, 34] uses the known region to replace the unmasked region in the diffusion process. Additionally, Stable Diffusion Inpainting [6] introduces random masking during the text-to-image (T2I) process for training, augmented by supplementary textual inputs for precise control. Smartbrush [8] exhibits the capability to tailor image results by manipulating mask types, while HD-Painter [9] and PowerPaint [10] further refine the capabilities of SDI through additional training. BrushNet [11] stands out as a cutting-edge inpainting model, boasting plug-and-play functionality. Although these methods have yielded good results, there are still many difficulties in foreground-conditioned image inpainting.

2.3 Large Language Model for Vision Task

The field of natural language processing has experienced a dramatic transformation in the past time, with a record number of various large language model parameters and model capabilities approaching or even exceeding the human level [16, 18]. A number of high-performance models have also emerged in the field of visual question answering (VQA) [14, 15]. However, the high training costs have impeded the further advancement of visual language models. Leveraging existing large language models for visual tasks has become an important research direction [35], LLaVA [17], Bliva [36] made some attempts to align LLMs with visual features, and some researches [37, 38, 39, 40] employ LLMs as planners to assign downstream visual tasks based on different prompts. Woodpecker [41], SIRI [42] enhance the reasoning ability of VLMs through the knowledge of LLMs. There has been a trend to apply the capabilities of large models to multi-modality tasks.
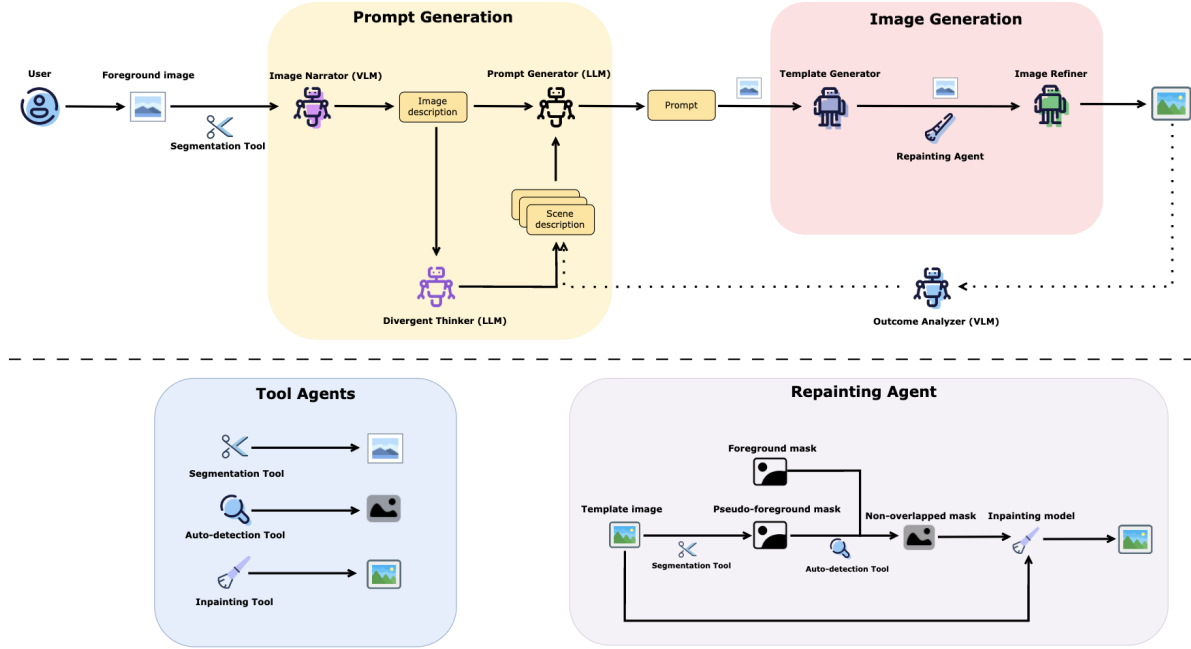
Figure 2: Illustration of our framework. Our framework consists of three modules, the prompt generation module, the image generation module and the outcome analyzer. The repainting agent processes auto-detection of "over-imagination" and re-inpainting the regions indicated by the non-overlapped mask with a text-guided image inpainting model.

3 Method

Anywhere is a multi-agent image generation framework comprising agents of various modalities, such as large language model, visual language model, controlled image generation model, and inpainting model. Its workflow encompasses three modules: the prompt generation module, the image generation module, and the outcome analyzer, as illustrated in Figure 2. Anywhere achieves background generation by processing images through modules utilizing diverse agents.

3.1 Prompt Generation Module

The prompt generation module aims to comprehend and associate the foreground to derive the background prompt. Firstly, the image narrator, embodied by a Vision-Language Model (VLM) agent, provides a textual description of the foreground's appearance attributes, encompassing color, texture, type, and viewpoint. We uphold a list of inquiries used as prompts of the VLM to gather valuable insights about the foreground objects. Secondly, the divergent thinker, portrayed by a Large Language Model (LLM), acts as a creative brainstormer, envisioning potential scenes in which the foreground could be situated based on the provided description. It generates a set of scene descriptions relevant to the foreground. We curate a repository of prompt templates for efficient brainstorming with the LLM.

Following this, the Prompt Generator, represented by LLM, assesses the relevance between scene descriptions and foreground descriptions, ranking the compatibility likelihood between the scene and foreground description. Ultimately, it selects the top-ranked scene description as the prompt. Furthermore, the type and viewpoint words of the foreground are integrated into the prompt as the final prompt. The process of prompt generation is outlined in Algorithm 1.

Algorithm 1 Prompt generation

Require: Foreground image i, vision language model $f_{\mathcal{V}}(x)$ (Image Narrator), large language model $f_{\mathcal{M}_1}(x)$ (Divergent Thinker), $f_{\mathcal{M}_2}(x)$ (Prompt Generator)
1: $t_i \leftarrow f_{\mathcal{V}}(i)$ {get foreground description}
2: $S \leftarrow f_{\mathcal{M}_1(t_i)}$ {get a set of scene descriptions}
3: $P \leftarrow f_{\mathcal{M}_2}(t_i, S)$ {determine the rank based on the relevance score between the foreground and scene description. }
4: prompt$\leftarrow f_{Top1}(P)$ {select top-ranked scene description}
5: **return** prompt

3.2 Image Generation Module

This module takes the foreground image and the prompt as inputs to generate a suitable scene for the foreground image. The template generator, implemented by the text-guided canny-to-image diffusion model, creates a scene image (template image) based on the prompt in a foreground-conditioned manner. Typically, the template image includes a similar subject (pseudo-foreground) to the foreground, serving as a mapping of the foreground in the scene.

Subsequently, the template image undergoes processing through the repainting agent, which inpaints the extraneous content surrounding the foreground, ensuring harmony between the foreground and template images after synthesis. The template image is segmented to obtain the pseudo-foreground, representing the foreground image in the new scene under the edge map condition. Usually, the pseudo-foreground image and the input foreground image do not perfectly overlap, and the area not covered will be repainted with the template image as input for the inpainting model. This process is illustrated in Figure 2.

However, the result from the repainting agent, after the copy-and-paste operation of the input foreground image, often exhibits blurring issues and edge artifacts, making it unsuitable as a final outcome. Thus, the image refiner, operated by the image-to-image diffusion model, corrects imperfections in the composite image, such as color discrepancies, shadow inconsistencies, or resolution adjustments.

Notably, our pipeline-based design of the image generation module prioritizes the use of an end-to-end text-guided image inpainting model, resulting in higher-quality outcomes and improved mitigation of "over-imagination".

3.3 Tool Agents

The tool agents refer to the tools or models employed in the framework, which target on various tasks. Our framework encompasses three types of tools, each serving specific responsibilities (as depicted in Figure 2). The segmentation tool is tasked with segmenting the foreground image, serving a dual role. Firstly, during the prompt generation module, it removes the background from the foreground image, resulting in a foreground-only image. Secondly, during the image generation module, it aids in comparing the foreground image with the pseudo-foreground segmented from the template image. The auto-detection tool is utilized to identify extraneous content surrounding the foreground in the template image. It achieves this by assessing the overlap between the pseudo-mask of the template image and the mask of the input foreground image. Ideally, complete overlap between the two masks indicates the absence of "over-imagination". The image inpainting model is then applied to restore the non-overlapping regions of the template image, when needed.

3.4 Outcome Analyzer

The outcome analyzer, operated by VLM, analyzes the outcomes of the image generation module, offering feedback for the next iteration. It assesses perspective consistency, foreground-background relevance, aesthetic score, and image content rationality. We curate a list of pertinent questions to serve as prompts for soliciting valuable feedback. This facilitates a comprehensive analysis of the outcomes. This feedback acts as the foundation for iterative refinement, with the divergent thinker incorporating feedback from previous rounds to enhance scene associations. This iterative feedback mechanism progressively improves the quality of prompts, thereby impacting the final outcome.

4 Experiments

4.1 Setup

Settings: In our framework, we use Gemini-Pro [43] as the LLM and Gemini-Pro-Vision [43] as the VLM. We chooses RMBG-1.4[2] as the segmentation tool, and LaMa [44] is also available. We utilize ControlNet_sdxl_canny[3] as the template generator and SDXL_inpainting[4] as the inpainting model used in the repainting agent. We choose SDXL refiner[5] as the image refiner.

Dataset:To assess our framework, we gathered photos of various entities from the internet, including common entities like cats, dogs, cars, boats, shoes, people, books, watches, etc. We prioritized selecting images with clear and distinct foregrounds to avoid incorrect foreground segmentation. Ultimately, we collected foreground images of 25 entities for experimentation. For the open-source model, we utilized these 25 foreground images to generate 4 results per foreground,

---

[2]https://huggingface.co/briaai/RMBG-1.4

[3]https://huggingface.co/diffusers/controlnet-canny-sdxl-1.0

[4]https://huggingface.co/diffusers/stable-diffusion-xl-1.0-inpainting-0.1

[5]https://huggingface.co/stabilityai/stable-diffusion-xl-refiner-1.0

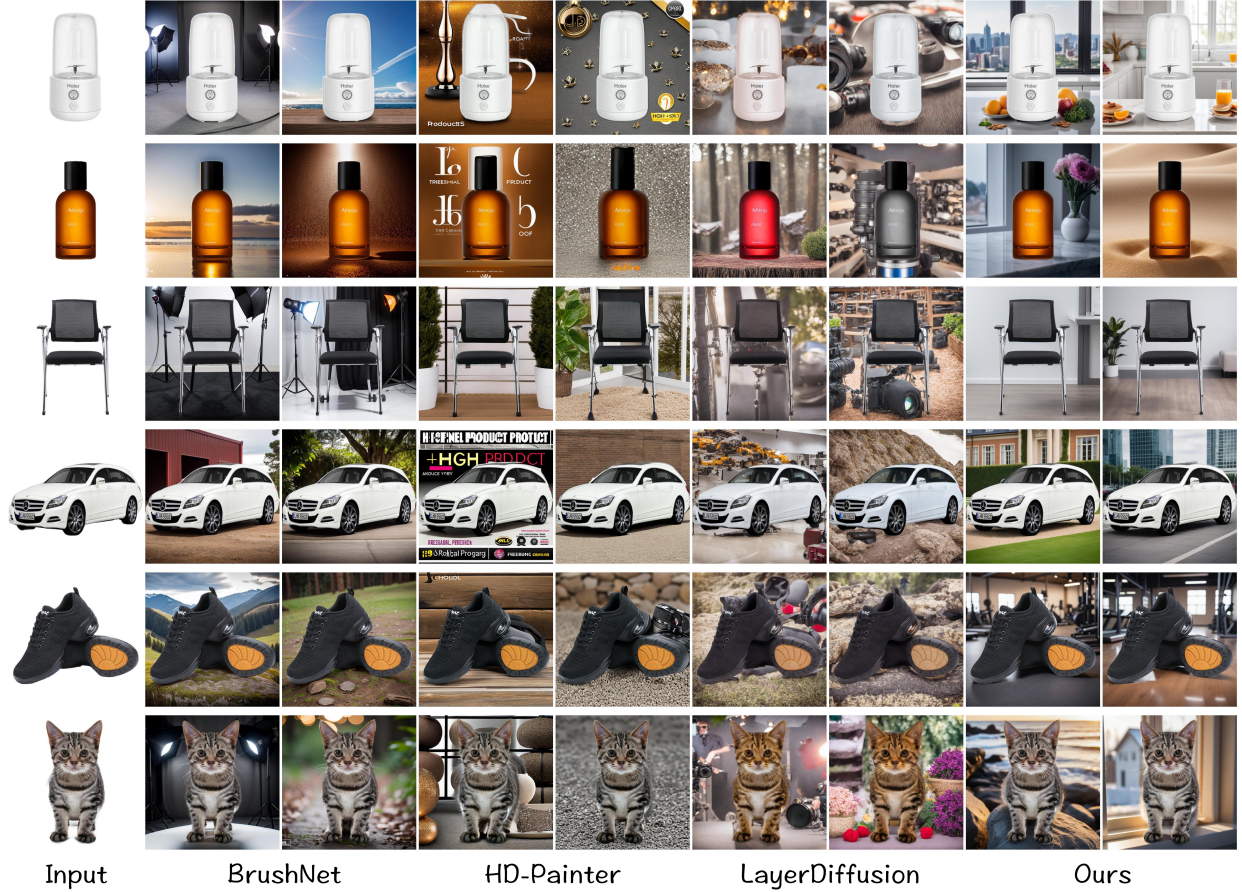|				
Input	BrushNet	HD-Painter	LayerDiffusion	Ours

Figure 3: Comparisons of our approach with the state-of-the-art open-source inpainting models. As shown, HD-painter and LayerDiffusion tend to produce results with artifacts or irrational contents: see Rows 2-6. BrushNet yields relatively satisfactory outcomes but lacks diversity, which can be observed from the frequent occurrence of photographic studios: see Rows 1, 3, and 6. In contrast, our methods produce high-quality and diverse results.

totaling 100 result images. For the commercial model, we uploaded the foreground images to the corresponding website to generate 2 results per foreground, resulting in a total of 50 results.

Baseline: We compared our method with current SOTA inpainting models: BrushNet [11], HD-Painter [9], LayerDiffusion [12]. Additionally, for a broader comparison, we experimented with some commercial products, including Phot.ai[6], Mokker.ai[7], Flair.ai[8].

Metrics: To assess the quality of the results, we established three metrics: aesthetic score, diversity score, and bad case rate.

- The aesthetic score evaluates the satisfaction level of the results on a scale of 1 to 5 points. A score of 1 indicates multiple evident issues in the image, such as foreground-background inconsistency or evident "over-imagination". A score of 2 indicates one noticeable issue in the image, while 3 indicates no significant issues overall but minor flaws in details, like unnatural lightening, shadowing, or edge boundaries, or slight "over-imagination". A score of 4 suggests good quality suitable for display, and 5 denotes an exceptionally visually compelling image with astonishing effects.

- The diversity score gauges the variety of results generated while maintaining consistency between foreground and background. It ranges from 1 to 3 points, with 1 indicating a single scene generated in the results, such as monochrome or similar scenes. A score of 2 indicates some similarity among scenes in the results, while 3 signifies that all scenes in the results are highly novel.

---

[6]https://www.phot.ai/
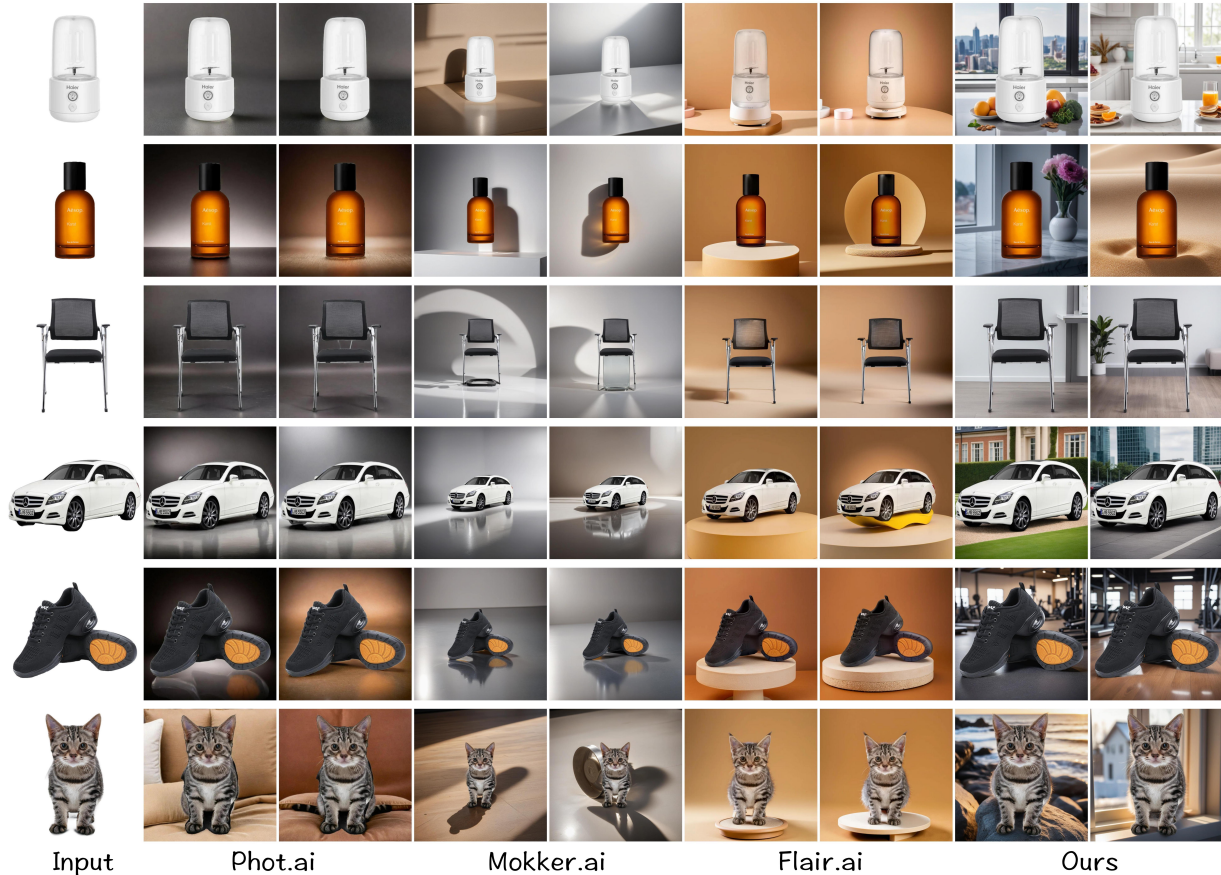[7]https://app.mokker.ai/
[8]https://app.flair.ai/

Figure 4: Comparisons between our framework with pioneering commercial systems. As shown, while commercial systems often prioritize avoiding poor generation results at the expense of diversity, our approach ensures both reliability and diversity.

- The bad case ratio measures the usability of the results. Each result is assessed to determine whether it is a bad case, marked as yes or no. If there are one or more evident issues in the result, it is considered a bad case.

4.2 Qualitative Result

Comparison results with open-source models are depicted in Figure 3. Our framework excels in several aspects: it produces backgrounds more suitable for foregrounds, offers diversity in background generation, and effectively addresses "over-imagination" issues during the inpainting process. In the third rows of Figure 3, other methods exhibit instances of "over-imagination" with chairs displaying additional "legs" or extra components. Our method successfully avoids such occurrences. In the first rows of Figure 3, while other approaches struggle to generate relevant background scenes for the kitchen blender, our framework adeptly comprehends the kitchen scenario.

Comparison results with commercial products are displayed in Figure 4. For these commercial systems, only foreground image information is provided. It is evident from the results that these commercial products can only generate a single background under a given template scene, and instances of mental imagery phenomena are observed. For instance, as illustrated in Columns 6 and 7 of the blender rows in Figure 4, legs or stands are erroneously added to the blender. Similarly, in Columns 3, 4, and 5 of the chair rows, unwanted accessories are imposed upon the chairs. In contrast, our framework excels in generating imaginative backgrounds while adapting to diverse foreground types and preserving foreground integrity.

4.3 Quantitative Result

We utilize human evaluators to assess the generation results based on the aforementioned standards. Subsequently, we calculate the average scores for aesthetic and diversity across all test cases. The bad case rate is determined by tallying
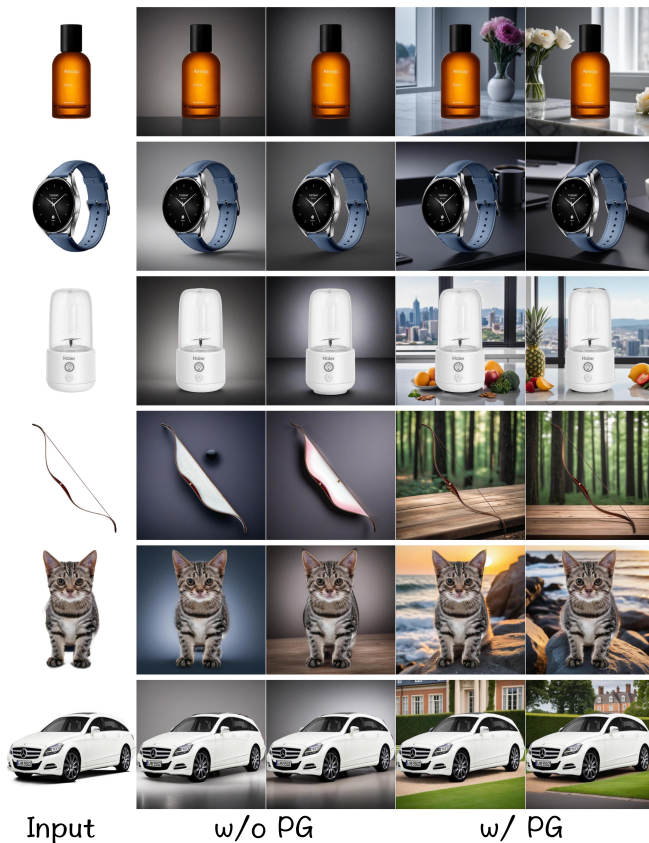
Figure 5: Ablation studies on the prompt generation module (PG). It reveals that without the prompt generation module, our system tends to produce less diverse results with uniform empty backgrounds.

Method	Aesthetic score↑	Diversity score ↑	bad case rate ↓
BrushNet	2.98	2.36	0.33
HD-Painter	1.91	2.12	0.64
LayerDiffusion	1.93	1.96	0.50
Phot.ai	3.05	1.20	0.30
Mokker.ai	3.40	1.55	0.34
Flair.ai	3.32	1.25	0.20
Ours	**3.52**	**2.52**	**0.18**

Table 1: Quantitative analysis of our framework, state-of-the-art open-source inpainting models, and pioneering commercial systems. It indicates that while the selected commercial systems generate results with less diversity, they exhibit relatively high reliability and quality. Among all the open-source models, BrushNet performs the best on average but demonstrates lower diversity and quality compared to ours.

the occurrences of bad cases among all generation samples. The quantitative results on the dataset are presented in Table 1.

As shown in Table 1, the aesthetic score and diversity score achived with our framework outperform those of both open-source models and commercial products, while also achieving the lowest bad case rate. Notably, commercial products exhibit lower scores in diversity, possibly attributed to their utilization of fixed templates for reliable results, albeit at the expense of diversity. On the other hand, open-source models can offer relatively high diversity but often yield unreliable results. Our method strikes a balance by delivering diversified results while ensuring reliability.
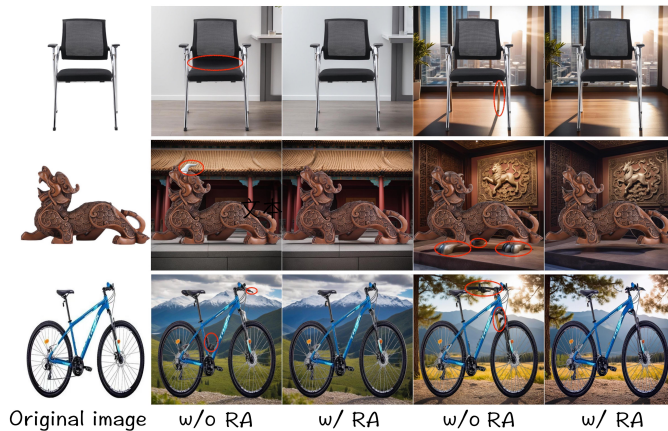
Figure 6: Ablation studies on the repainting agent (RA). The red circles highlight the regions with "over-imagination". As shown, the repainting agent contributes to mitigating the "over-imagination" issue.

4.4 Ablation Study

To assess the effectiveness of the design of our framework, we conducted ablation studies on three different modules and functionalities, the prompt generation module, the repainting module, and the outcome analyzer.

Prompt Generation Module

To assess the impact of removing the prompt generation module, we deactivate the module and provide the image generation module with a generic description, such as "a photograph" or "an imaginary scene". Figure 5 showcases the results with and without the prompt generation module, which indicates the impact of removing the prompt generation module. We can conclude from Figure 5 that the prompt generation module not only introduces diversity into the results but also helps mitigate the occurrence of "over-imagination" to some extent: see the bow in the fourth rows of Figure 5.

Repainting Agent

To evaluate the significance of the repainting agent, we exclude it from the process and directly feed the template image to the image refiner. Figure 6 illustrates the results with and without the repainting agent. As shown, results without the repainting agent may exhibit "over-imagination", whereas this tool effectively addresses inconsistencies arising from such occurrences.
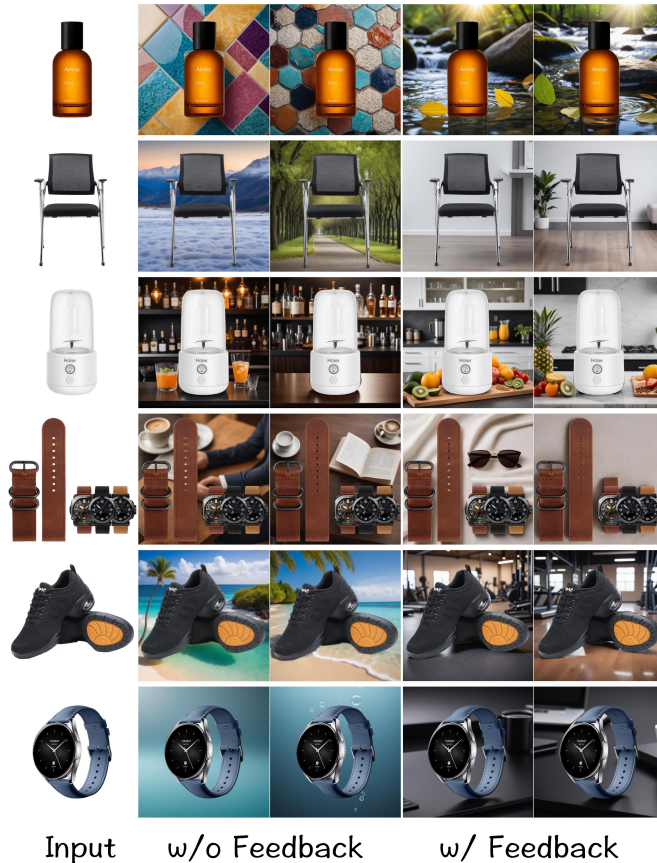
Input   w/o Feedback   w/ Feedback

Figure 7: Ablation studies on the use of the outcome analyzer. As shown, the mechanism of feedback-based regeneration significantly improves the quality of the final outcomes. Rows 1 and 4 indicates instances of view inconsistency without the feedback-loop. Rows 2, 3 and 6 indicates foreground-background irrelevance without the feedback-loop. Rows 5 indicates content irrationality due to erroneous relative size without the regeneration mechanism.

Feedback Mechanism

To assess the impact of the outcome analyzer's feedback loop, we'll bypass it and directly evaluate the raw output of the image generation module in a single pass. Figure 7 demonstrates the results with and without the feedback mechanism of the outcome analyzer. As shown, initially, issues like foreground-background inconsistency or improper viewpoints may be present. However, results after iterative feedback from the outcome analyzer learn from previous problems and yield improved outcomes.

5 Conclusion and Future Work

In this paper, we introduce a novel multi-agent framework for robust and diverse foreground-conditioned image inpainting. Our method demonstrates a 12% decrease in bad case rate, along with a 0.16 increase in diversity score and a 0.54 increase in aesthetic score compared to the leading state-of-the-art approach. Additionally, our framework achieves a 2% reduction in bad case rate, with a 0.12 increase in aesthetic score and a 0.97 increase in diversity score compared to the pioneering commercial system, representing a significant advancement in the field of foreground-conditioned image inpainting.

However, our approach faces certain limitations. Firstly, it struggles with foreground objects containing transparent or semi-transparent components (e.g., glass cup, magnifier). Secondly, the outcome analyzer encounters challenges in predicting image rationality related to lighting and shadowing, leading to some unsatisfactory outcomes. With advancements in VLM, LLM, and image generators, the results of our framework could be further improved. Meanwhile, we will attempt to enhance our approach's capabilities in addressing these challenges by optimizing the pipeline design in future research.

References

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

[2] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[3] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023.

[4] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18370–18380, 2023.

[5] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.

[6] 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.

[7] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[8] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023.

[9] Hayk Manukyan, Andranik Sargsyan, Barsegh Atanyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Hd-painter: High-resolution and prompt-faithful text-guided image inpainting with diffusion models. *arXiv preprint arXiv:2312.14091*, 2023.

[10] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. *arXiv preprint arXiv:2312.03594*, 2023.

[11] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. *arXiv preprint arXiv:2403.06976*, 2024.

[12] Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *arXiv preprint arXiv:2402.17113*, 2024.

[13] Yongsheng Yu, Hao Wang, Tiejian Luo, Heng Fan, and Libo Zhang. Magic: Multi-modality guided image completion. *arXiv preprint arXiv:2305.11818*, 2023.

[14] 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.

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[16] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[18] 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.

[19] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[20] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[21] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation–supplementary materials–.

[22] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023.

[24] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023.

[25] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023.

[26] Michelle Shu, Charles Herrmann, Richard Strong Bowen, Forrester Cole, and Ramin Zabih. Dreamwalk: Style space exploration using diffusion guidance. *arXiv preprint arXiv:2404.03145*, 2024.

[27] Guillaume Couairon, Marlène Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2174–2183, 2023.

[28] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023.

[29] Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. *arXiv preprint arXiv:2312.01305*, 2023.

[30] Ziyang Yuan, Mingdeng Cao, Xintao Wang, Zhongang Qi, Chun Yuan, and Ying Shan. Customnet: Zero-shot object customization with variable-viewpoints in text-to-image diffusion models. *arXiv preprint arXiv:2310.19784*, 2023.

[31] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.

[32] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

[33] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.

[34] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42(4):1–11, 2023.

[35] 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.

[36] Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2256–2264, 2024.

[37] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

[38] Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv preprint arXiv:2306.08640*, 2023.

[39] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36, 2024.

[40] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898, 2023.

[41] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

[42] Zeqing Wang, Wentao Wan, Runmeng Chen, Qiqing Lao, Minjie Lang, and Keze Wang. Towards top-down reasoning: An explainable multi-agent approach for visual question answering. *arXiv preprint arXiv:2311.17331*, 2023.

[43] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[44] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022.

A   Detailed prompt template of LLM and VLM

There are different prompt template in our framework (see Figure 2).

The image narrator receives input image, which are processed by the following prompt template to get a image description, we use the following prompt template:

```
1  You are an analyst and observer, you can give a detailed description of any object
       and discover the characteristics of that object. Please give a detail description
       of this image, as well as describing the important features in that image, and
       then give the name and the viewpoint of this object. Please provide a response in
       a structured JSON format that matches the following model: {YOUR_JSON_FORMAT}.
```

The divergent thinker receives image description for scene association and gets a candidate set of scene (size N). The prompt template is as follow:

```
1  You are an expert imaginative photographer, you can choose a variety of suitable
       scenes for any object. I'll asking you to provide me with scene descriptions,
       then I'll provide some useful information for you: the object infomation: [{
       object_name}], the viewpoint: [{viewpoint}] must appear in scene description, and
        feedback about the object's previous scene result is: [{feedback}]. Please give
       5 sets of relevant scene descriptions for this object: [{prompt}]. Please provide
        a response in a structured JSON format that matches the following model: {
       YOUR_JSON_FORMAT}.
```

The prompt generation receives the image descriptions and candidate set of scene for sorting, which is used to get the best scene description, the prompt template is:

```
1  You are an excellent analyst, able to see the correlation between different texts.
       Now we have a object description:[{img_desc}]. Please give me the sort number (
       from 1 to 5) about these 5 scene description: [{scene_descs}] that most
       appropriate with the object. Please provide a response in a structured JSON
       format that matches the following model: {json_format}.
```

The outcome analyzer analyze of the problems in the resultant image is used to feed back into the divergent thinker so that the unconsistency results of the previous round can be considered in the next round, we use the following prompt template:

```
1  You are an analyst expert and an observer of detail. Please give the answer of these
       questions: "Is it common for the [{subject}] to be placed in this context?" , Is
       [{subject}] placed normally on a platform or on the ground?. Please provide a
       response in a structured JSON format that matches the following model: {
       json_format}.
```

B   More qualitative results

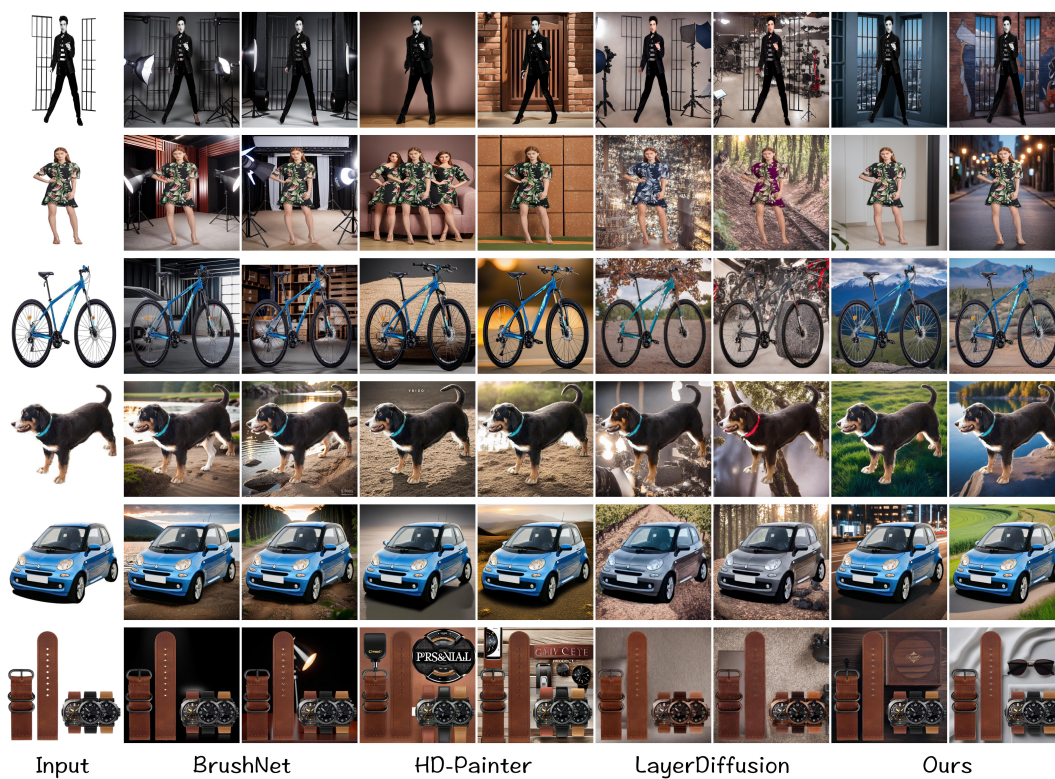There are more comparisons of SOTA open-source inpainting models in Figure A1 and Figure A2.
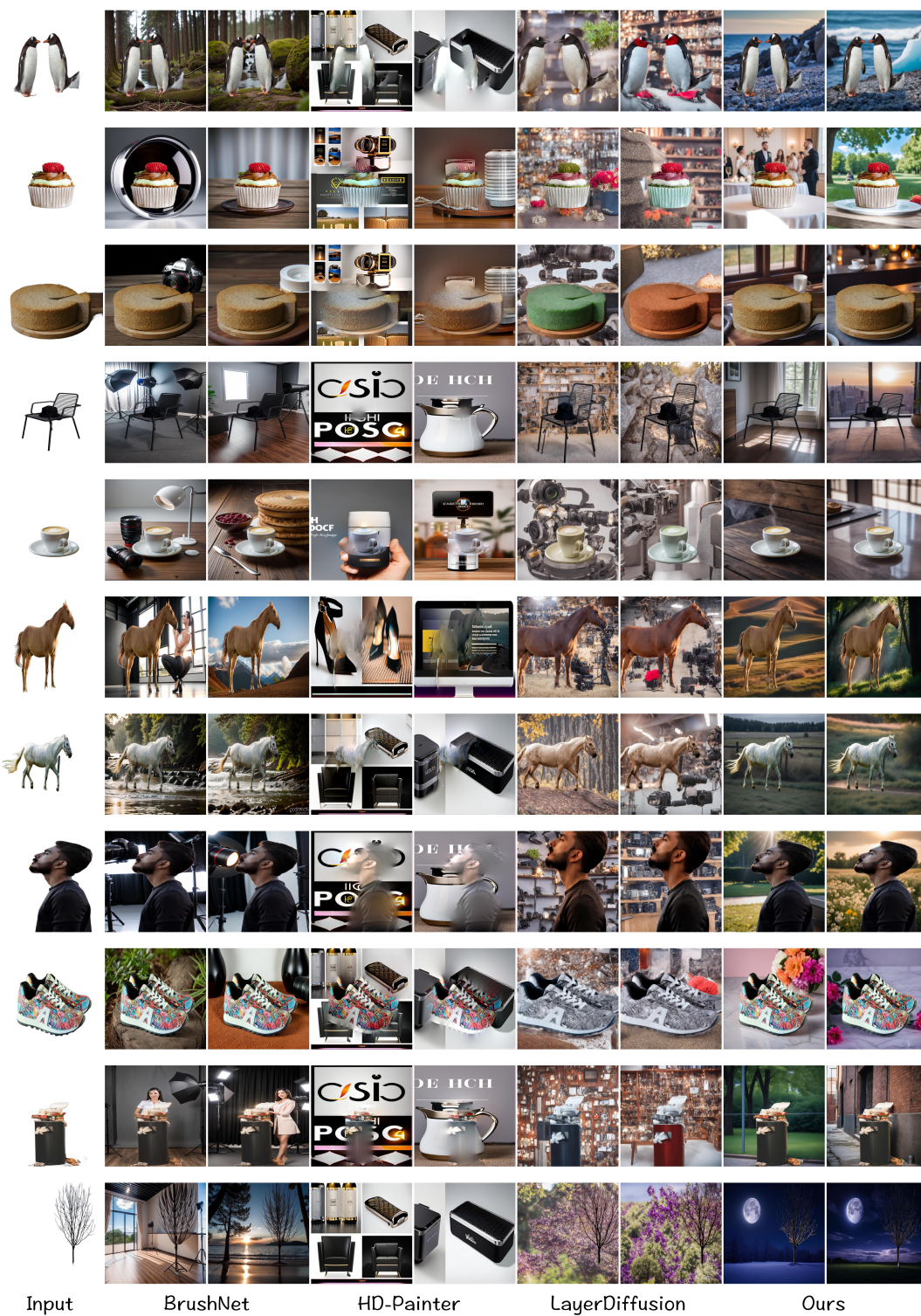
Figure A1: Comparisons of SOTA inpanting model.

Figure A2: Comparisons of SOTA inpanting model.