Anywhere: A Multi-Agent Framework for User-Guided, Reliable, and Diverse Foreground-Conditioned Image Generation

Xie Tianyidan[2], Rui Ma[3], Qian Wang[4*], Xiaoqian Ye[4], Feixuan Liu[5], Ying Tai[1,2], Zhenyu Zhang[1,2], Lanjun Wang[6], Zili Yi[1,2]*

[1]State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing, China
[2]School of Intelligence Science and Technology, Nanjing University, Suzhou, China
[3]Jilin University, Changchun, China
[4]China Mobile Research Institute, Beijing, China
[5]Beijing Shuzhimei Technology Co., Ltd, Beijing, China
[6]School of New Media and Communication, Tianjin University, Tianjin, China
sealical@outlook.com, yi@nju.edu.cn

Abstract

Recent advancements in image-conditioned image generation have demonstrated substantial progress. However, foreground-conditioned image generation remains underexplored, encountering challenges such as compromised object integrity, foreground-background inconsistencies, limited diversity, and reduced control flexibility. These challenges arise from current end-to-end inpainting models, which suffer from inaccurate training masks, limited foreground semantic understanding, data distribution biases, and inherent interference between visual and textual prompts. To overcome these limitations, we present Anywhere, a multi-agent framework that departs from the traditional end-to-end approach. In this framework, each agent is specialized in a distinct aspect, such as foreground understanding, diversity enhancement, object integrity protection, and textual prompt consistency. Our framework is further enhanced with the ability to incorporate optional user textual inputs, perform automated quality assessments, and initiate re-generation as needed. Comprehensive experiments demonstrate that this modular design effectively overcomes the limitations of existing end-to-end models, resulting in higher fidelity, quality, diversity and controllability in foreground-conditioned image generation. Additionally, the Anywhere framework is extensible, allowing it to benefit from future advancements in each individual agent.

1 Introduction

Image generation conditioned on visual inputs has made remarkable strides in recent years, fueled by advancements in diffusion models (Ho, Jain, and Abbeel 2020; Zhang, Rao, and Agrawala 2023; Huang et al. 2023; Avrahami et al. 2023; Li et al. 2023). These models have enabled sophisticated techniques for tasks such as inpainting, image expansion, and object insertion (Rombach et al. 2022; Podell et al. 2023; Manukyan et al. 2023; Ju et al. 2024; Zhang and Agrawala 2024). However, the image generation task that attempts to complete the background based on a given foreground object remains underexplored. This technique enhances content creation, e-commerce visualization, and gaming by generating contextually appropriate backgrounds. Its significance lies in its wide-ranging applications, including virtual try-on, personalized advertising, and augmented reality.

The complexity of foreground-conditioned image generation stems from its multifaceted nature, requiring simultaneous attention to object integrity, contextual relevance, and creative diversity. This task demands a deep understanding of visual semantics, spatial relationships, and creative composition. Existing inpainting methods often fail in three critical aspects: see Fig. 1.

- **Violated object integrity**: Existing inpainting methods often struggle to maintain the integrity of foreground objects, leading to the generation of unwanted elements or extensions. This issue arises because these methods rely on a single network to generate backgrounds in an end-to-end manner, heavily dependent on large datasets of foreground-background pairs obtained through auto-labeling with existing segmentation models or random masking. However, the accuracy of these segmentation masks is not always reliable, resulting in compromised object integrity and the inadvertent introduction of unwanted elements around the foreground object.
- **Foreground-background inconsistency**: Current image generation models, through trained to extend coherent structures from the input foreground, lack a deep semantic understanding of the foreground object and its relationship to the background, leading to contextually inappropriate or implausible scenes.
- **Limited diversity**: Existing models tend to generate monotonous or stereotypical backgrounds, failing to fully explore creative possibilities, as existing image generation models tend to incorporate and amplify biases from their training data (e.g., photographs with uniform or monotonous backgrounds).
- **Compromised Textual Consistency**: When applied to text-guided inpainting, the consistency of textual prompts can be compromised due to mutual interference that occurs during the joint integration of visual and textual inputs. This issue arises because text-guided inpaint-

*Corresponding author

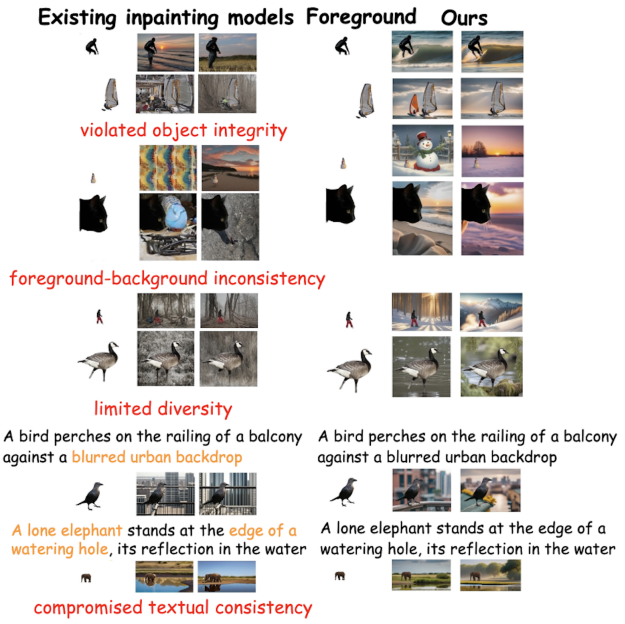

Figure 1: Comparison of our Anywhere framework with inpainting models for foreground-conditioned image generation. The left section highlights the limitations of existing inpainting models, while the right section showcases our results. Our approach effectively addresses the issues (e.g., violated object integrity, foreground-background inconsistency, limited diversity, and compromised textual consistency), producing foreground-preserved, semantically coherent, diverse and text-consistent backgrounds tailored to the given foreground objects.

ing models are typically adapted from text-to-image generators, which lack effective mechanisms and sufficient supervision to prevent unhealthy mutual interference between the visual and textual conditions.

Recognizing the limitations of end-to-end models, we propose a modular approach that incorporates multiple specialized agents to address the problem. Specifically, we introduce the Foreground Analyzer, based on advanced Visual Language Models (VLM) (Alayrac et al. 2022; Li et al. 2022; Achiam et al. 2023; Liu et al. 2024), to achieve a deep understanding of foreground semantics. To enhance the diversity of generated images, the Prompt Creator leverages Large Language Models (LLM) (Touvron et al. 2023; Achiam et al. 2023) to produce creative textual prompts. In particular, we design the Template Repainter to protect object integrity by automatically detecting violations of object integrity and initiating repainting when necessary. Furthermore, the Quality Analyzer, also based on VLM, performs automated quality assessments and triggers regeneration if needed. In addition, our framework is extended to allow optional user textual inputs, composing prompts by merging these inputs with foreground semantics.

We conducted comprehensive evaluations of our framework, demonstrating that the Anywhere framework significantly reduces instances of violated object integrity, improves foreground-background consistency, enhances diversity and controllability in foreground-conditioned image generation. Both subjective and objective metrics confirm that our approach excels in quality, diversity, user preference, and user controllability. In summary, the major contributions of this paper include:

- We introduce an innovative multi-agent framework specifically designed for foreground-conditioned image generation, representing a significant departure from traditional end-to-end models. This approach effectively overcomes the limitations of existing methods, leading to substantial improvements in the quality, robustness, diversity, and controllability of the generated results.
- We design a Template Repainting Agent equipped with a unique mechanism for preserving object integrity and adaptive background synthesis. This agent successfully mitigates issues related to object integrity while maintaining contextual relevance, as validated by large-scale experiments.
- Extensive evaluations reveal that our framework achieves a 4.6% improvements in FID and an average 24% increase in user preference scores, reduce 44% of bad cases, along with a 33% boost in the diversity score, compared to the best state-of-the-art inpainting models. In scenarios involving user textual inputs, our framework demonstrates a 5% increase in text-image matching score over the leading text-guided image inpainting models.

2 Related Works

2.1 Diffusion-based Controllable Image Generation

Stable Diffusion, a leading text-to-image (T2I) model, has rapidly evolved beyond simple text inputs. While some researchers explore text-driven image-to-image generation (Hertz et al. 2022; Brooks, Holynski, and Efros 2023), others have introduced diverse control signals to enhance the diffusion process. These include subject images (Gal et al. 2022; Ruiz et al. 2023), style information (Sohn et al. 2023), layout conditions (Avrahami et al. 2023), edge maps (Zhang, Rao, and Agrawala 2023), segmentation masks (Couairon et al. 2023), and viewpoint control (Liu et al. 2023a). Notably, LayerDiffusion (Zhang and Agrawala 2024) generates images on transparent layers, allowing foreground or background elements to guide the process. These advancements demonstrate diffusion models' expanding capabilities to create more diverse and precise user-tailored images.

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 (Saharia et al. 2022) and Repaint (Lugmayr et al. 2022), which leverage the original image alongside the unmasked regions to enhance denoising. Blended Diffusion (Avrahami,

Lischinski, and Fried 2022; Avrahami, Fried, and Lischinski 2023) uses the known region to replace the unmasked region in the diffusion process. Additionally, Stable Diffusion Inpainting (Rombach et al. 2022) introduces random masking during the text-to-image (T2I) process for training, augmented by supplementary textual inputs for precise control. Smartbrush (Xie et al. 2023) exhibits the capability to tailor image results by manipulating mask types, while HD-Painter (Manukyan et al. 2023) and PowerPaint (Zhuang et al. 2023) further refine the capabilities of SDI through additional training. BrushNet (Ju et al. 2024) 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 generation, facing challenges such as violated object integrity where excessive content compromises foreground integrity, foreground-background inconsistency producing contextually inappropriate backgrounds, limited diversity and text-consistency in generated backgrounds. Hence more advanced approaches are needed for foreground-conditioned image generation.

2.3 Large Language Model for Vision Task

Natural language processing has undergone a dramatic transformation, with large language models (LLMs) approaching or surpassing human-level capabilities (Achiam et al. 2023; Touvron et al. 2023). Simultaneously, visual question answering (VQA) has seen the emergence of high-performance models (Alayrac et al. 2022; Li et al. 2022). Despite high training costs impeding visual language model advancement, leveraging existing LLMs for visual tasks has become a key research direction (Brown et al. 2020). Models like LLaVA (Liu et al. 2024) and Bliva (Hu et al. 2024) align LLMs with visual features, while others use LLMs as planners for visual tasks (Wu et al. 2023a; Gao et al. 2023; Shen et al. 2024; Surís, Menon, and Vondrick 2023). Woodpecker (Yin et al. 2023) and SIRI (Wang et al. 2023) enhance VLM reasoning through LLM knowledge. This trend reflects the growing application of large models to multi-modal tasks.

3 Method

3.1 Framework Overview

Anywhere is a multi-agent framework specifically designed to tackle the challenges of foreground-conditioned image generation. This framework integrates LLMs, VLMs, and image generation models into a sophisticated pipeline, as illustrated in Fig. 2 (a). The framework consists of three primary components: The Prompt Generation Module generates an elaborate textual prompt by leveraging the semantic understanding of the input foreground contents and the creative capabilities of LLMs; The Image Generation Module then utilizes the optimized textual prompt to create an appropriate template. The Quality Evaluator assesses the final image quality, providing descriptive and information-rich feedback to facilitate the re-generation of results.

3.2 Prompt Generation Module

The Prompt Generation Module employs specialized agents focused on foreground understanding and textual prompt optimization to address foreground-background inconsistencies, incorporate optional user inputs, enhance text prompt consistency, and boost diversity. The key agents in this module are:

Foreground Analyzer This VLM-based agent extracts detailed textual information from the foreground image, capturing rich attributes of the foreground object, such as object type, shape, color, pose, and material. A comprehensive list of template questions guides the extraction of these specific attributes. The agent outputs finely detailed textual information in a structured, JSON-formatted description.

Prompt Creator This LLM-based agent generates diverse textual descriptions based on extracted foreground details and, when available, integrates user input and descriptive feedback. Acting as a creative engine, the agent explores imaginative compositions to enhance output diversity. It utilizes a pre-designed prompting scheme that ensures generated texts are both varied and compatible with image generation models. The agent combines three distinct inputs: foreground textual descriptions, user input if provided, and textual feedback from the Quality Evaluator, as illustrated in Fig. 2. By merging information from these sources, the agent generates k candidate textual prompts.

Prompt Selector This LLM-based agent evaluates the prompts generated by the Prompt Creator, taking into account factors such as relevance to the foreground details and compatibility with image generation models. It ranks the prompts based on evaluation scores, with the top-ranked prompts being selected probabilistically.

The Prompt Generation Module is designed to create textual prompts that facilitate the subsequent template generation process, ensuring relevance to the foreground, visual diversity, and text consistency. Additional details about this module are provided in the Appendix.

3.3 Image Generation Module

The Image Generation Module represents a major advancement in foreground-conditioned image generation by transforming template prompts into visually compelling backgrounds that seamlessly integrate with foreground images. Our multi-step scheme effectively addresses key challenges, particularly object integrity and foreground-background consistency. This is achieved through the use of three specialized agents that provide unparalleled control and precision throughout the generation process:

Template Generator We utilize ControlNet (Zhang, Rao, and Agrawala 2023), an edge-guided image generation model, to create initial background templates. This innovative application of ControlNet ensures spatial coherence between the generated backgrounds and the foreground objects, maintaining fine-grained control while producing high-quality images. The generation process is grounded in both the edge map of the foreground image and the template

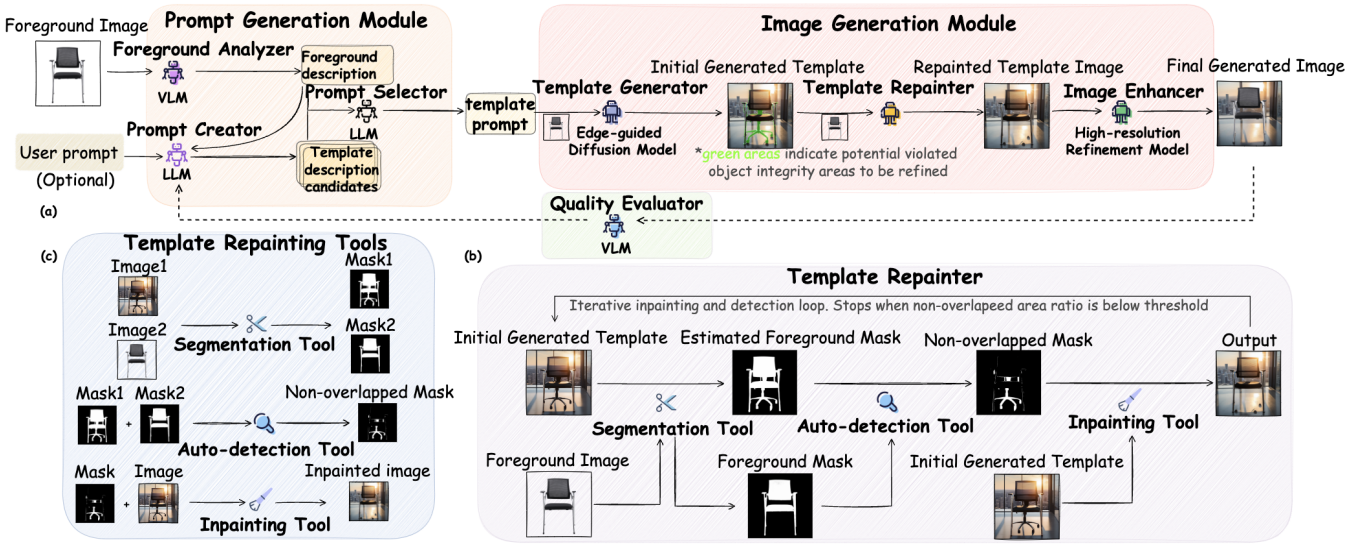

Figure 2: Overview of the Anywhere framework. (a) Our approach comprises three main components: the Prompt Generation Module, the Image Generation Module, and the Quality Evaluator (Agent). The Prompt Generation Module uses a Foreground Analyzer (VLM) to extract textual descriptions from the foreground and a Prompt Creator (LLM) to generate multiple textual prompts based on the foreground descriptions and the user textual inputs if provided. The multiple textual prompts are then assessed by the Prompt Selector (LLM) and the best matched prompt will be selected. The Image Generation module includes a Template Generator (edge-guided image generation model) that generates a template image based on the textual prompt, a Template Repainter that detects object integrity violations (highlighted in green) and resolves the violations if needed, and an Image Enhancer (high-resolution image refinement Model) to paste-back the foreground and harmonize the final output. The Quality Evaluator Agent (VLM) assesses the resulting image, providing descriptive feedback and triggering re-generation when needed. (b) Illustration of the Template Repainter that performs violation detection by foreground segmentation and mask contrasting, and inpaints violated regions if they exist. (c) Illustration of template repainting tools used in the framework.

prompt, establishing a solid foundation for contextually appropriate backgrounds.

Template Repainter This targeted agent allows for precise and efficient corrections, preserving object integrity and ensuring consistent foreground-background integration, as illustrated in Fig. 2 (b). It operates through the following key components (depicted in Fig. 2 (c)):

- **Segmentation Tool.** This tool generates an estimated foreground mask from the initial template image, which is further used by the auto-detection tool.
- **Auto-detection Tool.** This tool uses the estimated foreground mask and the actual foreground mask to detect areas where object integrity is compromised. It first identifies the foreground bounding box with detection model (Liu et al. 2023b) in the input image, then utilizes the bounding box to crop both masked images, and finally calculates a non-overlapped mask by comparing the cropped estimated and actual masks, pinpointing regions requiring repainting. Note that repainting is not triggered if the non-overlapped area is below a certain threshold.
- **Inpainting Tool**. This advanced model selectively inpaints the areas identified by the non-overlapped mask, generating contextually appropriate content.

Image Enhancer Powered by a high-resolution refinement model (such as Stable-Diffusion XL (Podell et al. 2023)), this agent enhances the overall quality of the composite image, focusing on fine details, color balance, and smooth transitions.

The collaborative effort of these agents transforms the template prompt and foreground image into a contextually appropriate and visually compelling final image. This process effectively realizes the creative vision established in the Prompt Generation Module while addressing the unique challenges of foreground-conditioned image generation. More detailed information about the Image Generation algorithm can be found in the Appendix.

3.4 Quality Evaluator

The Quality Evaluator, a VLM-based agent, enhances the final image quality through a feedback loop. This agent leverages the VLM's advanced capabilities to evaluate visual relationships, content rationality, and overall image quality, providing a nuanced and comprehensive assessment that surpasses traditional image quality metrics (Achiam et al. 2023; Team et al. 2023; Li et al. 2022). We have developed a detailed questionnaire prompt list to facilitate the analysis of foreground-background integration, focusing on factors such as lighting consistency, color harmony, structural coherence, spatial relationships, and semantic relevance.

Based on this analysis, the Quality Evaluator generates detailed textual feedback, which is communicated to the

Prompt Generation Module for potential re-generation. This feedback loop significantly enhances the system's ability to create reliable and high-fidelity compositions. To prevent endless generation iterations, the system is capped at a maximum of three loops, as validated by experimental results (see Appendix for more details).

4 Experiments

4.1 Experimental Setup

Dataset We curated a foreground dataset consisting of 3,000 images by randomly selecting 1,500 images from the LAION dataset (Schuhmann et al. 2022) and 1,500 from the MSCOCO dataset (Lin et al. 2014). Each original image was segmented, and a randomly chosen foreground segment was extracted to serve as the test data. This approach ensures a diverse range of in-the-wild scenarios and object types (e.g., humans, vehicles, pets).

Implementation Details Our framework integrates state-of-the-art models across its various components. The Prompt Generation Module leverages Gemini-Pro (LLM) (Team et al. 2023) for both the Prompt Creator and Prompt Selector, and Gemini-Pro-Vision (VLM) (Team et al. 2023) for the Foreground Analyzer. In the Image Generation Module, we employ ControlNet_SDXL_Canny (Diffusers 2023) as the Template Generator and SDXL Refiner (Stabilityai 2023) as the Image Enhancer. The Template Refinement Tools consist of RMBG-1.4 (BRIA 2024) for segmentation, Grounding DINO (Liu et al. 2023b) for auto-detection, and LaMa (Suvorov et al. 2022) for inpainting. Additionally, the Quality Evaluator uses Gemini-Pro-Vision (VLM) (Team et al. 2023). Detailed prompt templates for the LLM and VLM components are provided in the Appendix.

Baseline To assess the effectiveness of our proposed framework, we compared it with three state-of-the-art inpainting models: BrushNet (Ju et al. 2024), a plug-and-play dual-branch model for image inpainting; HD-Painter (Manukyan et al. 2023), a high-resolution inpainting model known for precise prompt adherence; and Stable Diffusion 2.0 Inpainting (Rombach et al. 2022; Stabilityai 2022), a widely-used diffusion-based inpainting model. These methods represent the cutting edge in image inpainting.

Evaluation Metrics We employ a comprehensive set of metrics to evaluate the aesthetic quality, human preference alignment, and image fidelity of the generated images. The Aesthetic Score (AS) (Schuhmann et al. 2022) assesses overall visual appeal, while PickScore (Kirstain et al. 2023) simulates human rating behavior. The Human Preference Score (HPS) (Wu et al. 2023b) directly measures human preference in comparison to real images, and the CLIP Similarity score (CLIP-Sim) (Radford et al. 2021) quantifies the semantic alignment between generated images and input text prompts, employing ViT-B/16 as its image encoder. Additionally, we use Fréchet Inception Distance (FID) (Heusel et al. 2017) to evaluate image naturalism and fidelity. This selection of metrics offers a thorough assessment of our framework's performance across aesthetic quality, human preference alignment, text-image coherence, image diversity, and fidelity.

4.2 Qualitative Results

The qualitative comparison results are presented in Fig. 3. In the text-free scenario, our approach generates contextually appropriate and diverse backgrounds while maintaining object integrity. In contrast, HD-Painter (HDP) and Stable Diffusion Inpainting (SDI) often produce inconsistent or illogical backgrounds and compromise object integrity. BrushNet (BN) tends to generate uniform backgrounds with limited creativity. In the text-guided scenario, our framework effectively integrates user text input to create coherent backgrounds that seamlessly blend with the foreground and prompt. HDP frequently violates object integrity (see rows 1-3 and 6) and overlooks key textual elements (see row 3). BN attempts to incorporate prompts but often places foregrounds in unsuitable settings (see rows 1-3 and 6), misses textual elements (see rows 2-4), and compromises object integrity (see row 4). SDI struggles to align its outputs with the given prompts, leading to irrelevant backgrounds (see rows 2-6).

4.3 Quantitative Results

We conducted quantitative experiments for both text-free (Image-to-Image, or I2I) and text-guided (Text-guided Image-to-Image, or TI2I) scenarios using the metrics described in Sec. 4.1. The comparative results are presented in Tab. 1.

Metric Calculation Details Our evaluation methodology generates one resulting image per model for each test case. The assessment process is structured as follows: Aesthetic Score (AS) is directly calculated from the images produced by each model, without the need for additional textual information. For metrics that require textual input (PickScore, HPS, IR), the text-free (I2I) scenarios use the foreground object type name extracted by the Foreground Analyzer as the textual input for evaluation. For instance, in Fig. 2, the foreground type is "chair". These metrics are then calculated for each generated image using the corresponding foreground label. In the text-guided (TI2I) scenarios, we conducted experiments with two types of user inputs: (1) generic phrases that cover both indoor and outdoor natural scenes, such as "sunset", "snow", "room", and "beach"; and (2) unique sentences tailored to each foreground image, generated using a VLM. The results presented for TI2I are the average scores across both the generic phrases and the VLM-generated unique sentences.

Results Analysis As shown, our framework consistently outperforms the state-of-the-art inpainting models across all metrics in both text-free (I2I) and text-guided (TI2I) scenarios. Our method achieves the highest scores in aesthetic quality and human preference, indicating superior visual appeal. In the text-guided scenario, our approach also shows significant improvements in text alignment, highlighting its effectiveness in enhancing textual consistency. While slight trade-offs are observed in some metrics, these reflect

Figure 3: We compare our approach to advanced inpainting models on foreground-conditioned image generation tasks in both text-free (I2I) and text-guided (TI2I) scenarios. These results are generated using unconstrained, in-the-wild foreground images. Red color indicates missing elements in generated images. The inpainting models used for comparison include HD-Painter (**HDP**), BrushNet (**BN**), and Stable Diffusion 2.0 Inpainting (**SDI**).

the inherent challenges of balancing foreground compatibility with prompt adherence. Notably, our framework consistently achieves the lowest FID scores across both tasks. These quantitative results align with our qualitative findings, confirming that the Anywhere multi-agent framework excels at generating visually appealing, diverse backgrounds while maintaining high relevance to both foreground objects and text prompts.

4.4 User Study

To validate our quantitative findings and assess real-world user preferences, we conducted a comprehensive user study involving 10 participants. The study evaluated a total of 100 foreground images, evenly divided between text-free (I2I) and text-guided (TI2I) scenarios for the foreground-conditioned image generation task. For the text-guided scenario, we randomly assigned prompts from a fixed list of textual templates to each image. Each test case was processed by our Anywhere framework and three state-of-the-art models: BrushNet (Ju et al. 2024), HD-Painter (Manukyan et al. 2023), and Stable Diffusion 2.0 Inpainting (Podell et al. 2023). To ensure a thorough evaluation, each comparative method generated 3 results per test case, resulting in a total of 1,200 images (100 test cases $\times$ 4 methods $\times$ 3 results per method). Participants were asked to assess three key aspects of the generated images: aesthetic quality (rated on a scale of 1-5, with 5 indicating the highest quality), diversity (rated on a scale of 1-3, with 3 indicating the highest level of diversity), and the identification of any bad cases they deemed unsatisfactory or problematic (e.g., object integrity violations,

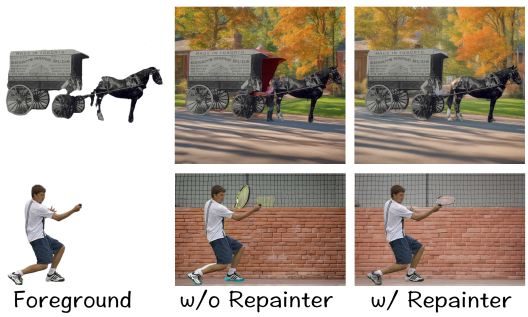

Figure 4: Ablation studies on the Template Repainter.

illogical content, severe artifacts, etc.). After collecting the ratings, The ratings for each method were averaged separately across the three evaluated aspects.

As shown in Tab. 2, our Anywhere framework consistently outperformed existing methods across all evaluated metrics. It achieved the highest aesthetic quality score, marking a significant improvement over the next best performer. The diversity score of our method substantially surpassed that of BrushNet, highlighting our framework's ability to generate a broader range of creative and contextually appropriate backgrounds. Additionally, our method exhibited the lowest bad case rate, demonstrating considerable improvements over both BrushNet and Stable Diffusion Inpainting. These results strongly validate the effectiveness of our framework in producing high-quality, diverse, and reliable outputs.

Task	Method	Aesthetic and Human Preference			Text Align	Image Fidelity
		AS (↑)	PickScore (↑)	HPS (↑)	CLIP-Sim (↑)	FID (↓)
I2I	BrushNet	5.238	0.261	0.186	-	41.047
	HD-Painter	4.899	0.133	0.178	-	38.797
	SD Inpainting	4.988	0.186	0.180	-	42.342
	Ours	**5.604**	**0.418**	**0.190**	-	**37.007**
TI2I	BrushNet	5.467	0.302	0.184	17.483	38.313
	HD-Painter	4.971	0.126	0.173	17.466	37.398
	SD Inpainting	5.132	0.177	0.175	17.314	39.476
	Ours	**5.812**	**0.394**	**0.195**	**18.357**	**36.450**

Table 1: Quantitative comparisons of our framework with advanced inpainting models on foreground-conditioned image generation tasks in both text-free (I2I) and text-guided (TI2I) scenarios.

Method	Aesthetic Quality (↑)	Diversity Score (↑)	Bad Case Rate (↓)
BrushNet	2.90	1.93	0.34
HD-Painter	1.86	1.78	0.66
SD Inpainting	2.12	1.85	0.47
Ours	**3.45**	**2.57**	**0.19**

Table 2: The user studies that compare our approach with advanced inpainting models by evaluating the aesthetics, diversity, and rate of bad cases in the generated results.

Task	Component Variant	AS (↑)	CLIP-Sim (↑)	FID (↓)
I2I	w/o PGM	5.263	-	40.751
	w/o TR	5.417	-	38.833
	w/o IE	5.518	-	37.980
	w/o FA	5.395	-	38.968
	w/o QE	5.462	-	38.510
	Ours	**5.604**	-	**37.007**
TI2I	w/o PC	5.561	18.077	37.346
	w/o PS	5.624	18.165	37.031
	w/o QE	5.537	18.211	36.952
	Ours	**5.812**	**18.357**	**36.450**

Table 3: Ablation studies for the Anywhere framework. PGM: Prompt Generation Module, TR: Template Repainter, IE: Image Enhancer, FA: Foreground Analyzer, QE: Quality Evaluator, PC: Prompt Creator, PS: Prompt Selector.

4.5 Ablation Study

For the ablation study, we systematically removed or modified various modules and assessed their impact on the evaluation metrics. In the ablation studies of the Prompt Generation Module, we directly set the user input text (if available) as the textual prompt for the Image Generation Module; if user input was not provided, an empty textual prompt was used instead. Without the Template Repainter, the initial template image was used as-is, without any additional processing. In the ablation studies of the Image Enhancer, we simply pasted the foreground object back onto the Template Repainter's output to produce the final result. To ablate the Foreground Analyzer, the foreground descriptions were excluded from the Prompt Creator's process. Without the Prompt Creator, we used the foreground descriptions concatenated with the user input text (if provided) as candidate template prompts for selection. For the Prompt Selector ablation, we randomly selected one of the outputs from the Prompt Creator as the template prompt.

Results in Tab. 3 demonstrate the impact of these ablation studies. The removal of the Prompt Generation Module resulted in significant performance drops in FID and aesthetics scores, underscoring its crucial role in enhancing quality and diversity. The absence of the Foreground Analyzer led to a greater decline in quality than the removal of the Prompt Creator or Prompt Selector, although the latter two are vital for maintaining textual consistency in text-guided scenarios. Both the Template Repainter and the Image Enhancer had a notable impact on image quality metrics. The Quality Evaluator, while contributing less to overall quality and text consistency, still played a role. The qualitative results in Fig. 4 highlight the importance of the Template Repainter in resolving issues of object integrity. Additional qualitative ablation results are presented in the Appendix.

5 Conclusion and Future Work

In this paper, we present Anywhere, a novel multi-agent framework for foreground-conditioned image generation that significantly outperforms existing end-to-end models in reliability, quality, diversity, and controllability. Our modular design, incorporating advanced VLMs, LLMs, and image generation models, effectively addresses critical challenges such as object integrity violations, foreground-background inconsistencies, limited diversity, and textual prompt inconsistencies. Our framework demonstrates substantial improvements over leading end-to-end inpainting models, with gains of 4.6% in FID, 24% in average human preference score, 33% in diversity score, and 5% in text-image matching score, while reducing bad cases by 44%.

However, these advancements come at the cost of increased computational demands, requiring approximately $2\sim3\times$ more GPU time than current end-to-end models on average for each generation. Additionally, the framework struggles with corner cases, such as transparent foreground objects. Future work will focus on optimizing computational efficiency and developing novel techniques to better handle long-tail scenarios.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62406134, No. 62202199) and the Nanjing University-China Mobile Communications Group Co. Ltd. Joint Institute.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Avrahami, O.; Fried, O.; and Lischinski, D. 2023. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42(4): 1–11.

Avrahami, O.; Hayes, T.; Gafni, O.; Gupta, S.; Taigman, Y.; Parikh, D.; Lischinski, D.; Fried, O.; and Yin, X. 2023. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18370–18380.

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18208–18218.

BRIA. 2024. BRIA Background Removal v1.4 Model Card.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Couairon, G.; Careil, M.; Cord, M.; Lathuilière, S.; and Verbeek, J. 2023. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2174–2183.

Diffusers. 2023. SDXL-controlnet: Canny.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gao, D.; Ji, L.; Zhou, L.; Lin, K. Q.; Chen, J.; Fan, Z.; and Shou, M. Z. 2023. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv preprint arXiv:2306.08640*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, W.; Xu, Y.; Li, Y.; Li, W.; Chen, Z.; and Tu, Z. 2024. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2256–2264.

Huang, L.; Chen, D.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2023. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*.

Ju, X.; Liu, X.; Wang, X.; Bian, Y.; Shan, Y.; and Xu, Q. 2024. BrushNet: A Plug-and-Play Image Inpainting Model with Decomposed Dual-Branch Diffusion. *arXiv preprint arXiv:2403.06976*.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22511–22521.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, R.; Wu, R.; Van Hoorick, B.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023a. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9298–9309.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11461–11471.

Manukyan, H.; Sargsyan, A.; Atanyan, B.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. HD-Painter:

High-Resolution and Prompt-Faithful Text-Guided Image Inpainting with Diffusion Models. *arXiv preprint arXiv:2312.14091*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. 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*, 22500–22510.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, 1–10.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; and Zhuang, Y. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

Sohn, K.; Ruiz, N.; Lee, K.; Chin, D. C.; Blok, I.; Chang, H.; Barber, J.; Jiang, L.; Entis, G.; Li, Y.; et al. 2023. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*.

Stabilityai. 2022. Stable Diffusion v2 Model Card.

Stabilityai. 2023. SD-XL 1.0-refiner Model Card.

Surís, D.; Menon, S.; and Vondrick, C. 2023. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11888–11898.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2149–2159.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, Z.; Wan, W.; Chen, R.; Lao, Q.; Lang, M.; and Wang, K. 2023. Towards Top-Down Reasoning: An Explainable Multi-Agent Approach for Visual Question Answering. *arXiv preprint arXiv:2311.17331*.

Wu, C.; Yin, S.; Qi, W.; Wang, X.; Tang, Z.; and Duan, N. 2023a. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023b. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*.

Xie, S.; Zhang, Z.; Lin, Z.; Hinz, T.; and Zhang, K. 2023. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22428–22437.

Yin, S.; Fu, C.; Zhao, S.; Xu, T.; Wang, H.; Sui, D.; Shen, Y.; Li, K.; Sun, X.; and Chen, E. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Zhang, L.; and Agrawala, M. 2024. Transparent Image Layer Diffusion using Latent Transparency. *arXiv preprint arXiv:2402.17113*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhuang, J.; Zeng, Y.; Liu, W.; Yuan, C.; and Chen, K. 2023. A Task is Worth One Word: Learning with Task Prompts for High-Quality Versatile Image Inpainting. *arXiv preprint arXiv:2312.03594*.