Yuan: Yielding Unblemished Aesthetics Through a Unified Network for Visual Imperfections Removal in Generated Images

Zhenyu Yu [1*], **Chee Seng Chan** [1]

[1] Faculty of Computer Science and Information Technology,
Universiti Malaya, Kuala Lumpur, 50603, Malaysia
yuzhenyuyxl@foxmail.com

Abstract

Generative AI presents transformative potential across various domains, from creative arts to scientific visualization. However, the utility of AI-generated imagery is often compromised by visual flaws, including anatomical inaccuracies, improper object placements, and misplaced textual elements. These imperfections pose significant challenges for practical applications. To overcome these limitations, we introduce *Yuan*, a novel framework that autonomously corrects visual imperfections in text-to-image synthesis. *Yuan* uniquely conditions on both the textual prompt and the segmented image, generating precise masks that identify areas in need of refinement without requiring manual intervention—a common constraint in previous methodologies. Following the automated masking process, an advanced inpainting module seamlessly integrates contextually coherent content into the identified regions, preserving the integrity and fidelity of the original image and associated text prompts. Through extensive experimentation on publicly available datasets such as ImageNet100 and Stanford Dogs, along with a custom-generated dataset, *Yuan* demonstrated superior performance in eliminating visual imperfections. Our approach consistently achieved higher scores in quantitative metrics, including NIQE, BRISQUE, and PI, alongside favorable qualitative evaluations. These results underscore *Yuan*'s potential to significantly enhance the quality and applicability of AI-generated images across diverse fields.

Code — https://github.com/YuZhenyuLindy/Yuan.git

Introduction

The field of generative artificial intelligence (AI) has witnessed substantial advancements, especially in text-to-image synthesis, as evidenced by recent studies (Li et al. 2023; Gafni et al. 2022; Gal et al. 2022). These technologies enable the creation of detailed and contextually accurate images from textual descriptions(Yao et al. 2010; Cetinic and She 2022; Wu et al. 2023). However, they often grapple with visual imperfections such as anatomical irregularities and inappropriate textual overlays, as depicted in Fig. 1. These flaws can significantly detract from the aesthetic and functional quality of the generated images.

Existing research often underestimates the importance of conditional removal of imperfections based on both text prompts and contextual image details. The ability to selectively edit or adjust elements within generated images with precision, guided by natural language, offers new opportunities in areas like interactive storytelling, bias mitigation, and ethical considerations. This approach allows users to refine visual content through linguistic cues, potentially challenging and reshaping biases within AI systems.

Current methods for addressing imperfections often rely on manual masks, which have several drawbacks: (i) they are labor-intensive and time-consuming, (ii) leading to inefficiency; (iii) their effectiveness is highly subjective and inconsistent, varying with individual skill; and they lack generalization, being limited to specific types of imperfections and not adaptable to diverse scenarios.

To address these challenges, we propose *Yuan*, a unified framework for automatically removing visual imperfections in text-to-image synthesis outputs. *Yuan* combines a grounded segmentation module, which identifies imperfections without predefined masks, and follow by an inpainting module that ensures contextually coherent restoration. Extensive experiments across diverse datasets demonstrate *Yuan*'s effectiveness, validating its efficiency in tasks such as image editing and content moderation. Case studies further highlight its practical utility, offering users greater control and flexibility in image manipulation.

In summary, this paper's contributions are:

- **Automated imperfection detection:** *Yuan* uses a novel segmentation module to automatically detect and outline visual imperfections, eliminating the need for manual masks and improving objectivity and consistency.
- **Context-aware inpainting:** The inpainting module seamlessly repairs identified imperfections, preserving the visual and contextual integrity of the images and enhancing their quality.
- **Comprehensive validation:** *Yuan* demonstrates superior performance across various datasets and scenarios, validated through quantitative metrics and qualitative assessments, proving its adaptability in diverse applications.

These contributions position *Yuan* as a scalable, user-friendly solution that sets new standards for automatic visual refinement in text-to-image synthesis.

*Corresponding author.

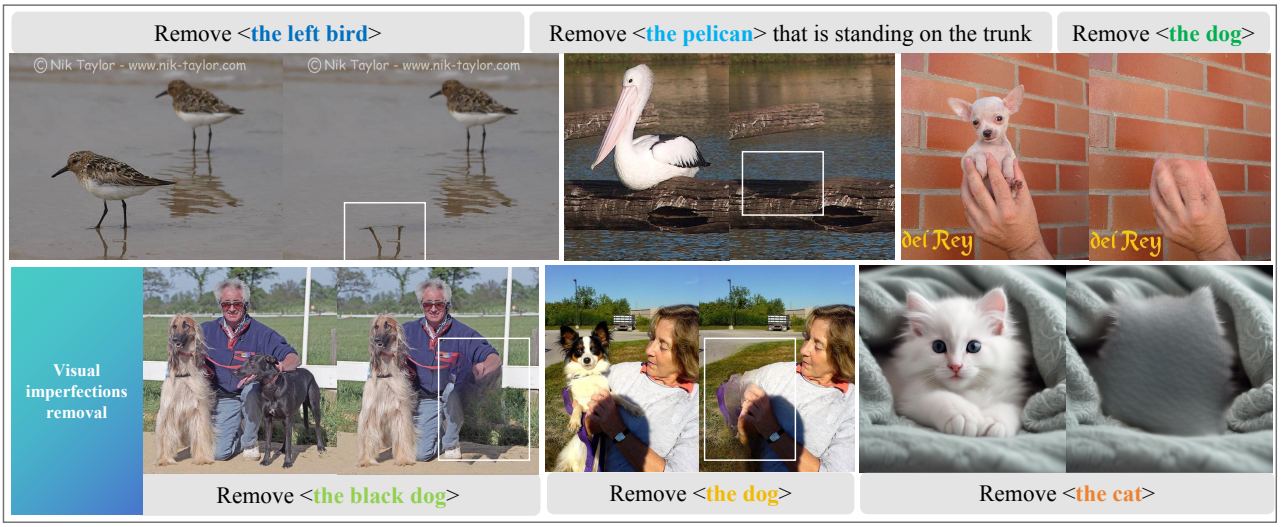

Figure 1: **Motivation for the study:** Existing algorithms for target removal often fall short in addressing related elements, such as reflections and shadows, resulting in incomplete or unnatural outcomes. Additionally, the removal of specified content can leave behind visual inconsistencies, such as unnatural postures or actions, necessitating further corrections. These challenges underscore the need for more advanced methods to achieve coherent and realistic image modifications.

Related Work

Image Generation and Synthesis

Image Generation Generative models have revolutionized the field of image synthesis, with significant contributions from models such as Generative Adversarial Networks (GANs) (Goodfellow et al. 2014; Yu and Wang 2024; Wang, Yang, and Yu 2024) and Variational Autoencoders (VAEs) (Kingma and Welling 2013). GANs, in particular, have achieved remarkable success in generating high-quality, realistic images through their dual-network structure, comprising a generator and a discriminator. Recent advancements, including StyleGAN (Karras, Laine, and Aila 2019; Yu et al. 2023) and BigGAN (Brock, Donahue, and Simonyan 2018), have further extended the resolution and diversity of generated images. Despite these successes, these models often produce outputs with visual imperfections, such as artifacts and inconsistencies, particularly in scenarios involving fine details or complex backgrounds.

Text-to-image Synthesis Text-to-image synthesis is a rapidly advancing field focused on generating images from textual descriptions, bridging the gap between natural language and visual content. Techniques like DALL-E (Ramesh et al. 2021) and Imagen (Saharia et al. 2022) use transformers and diffusion models to convert text prompts into detailed images. While these models can create intricate and contextually relevant visuals, they often face challenges such as anatomical inaccuracies and misplaced textual elements (Xu et al. 2018). Diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2020; Yu 2024), which generate high-quality images through a reverse denoising process, have shown great promise in improving the stability and diversity of image synthesis, leading to more realistic and detailed outputs. Combining diffusion models with text encoders, as seen in DALL-E 2 (Ramesh et al. 2022) and Imagen, has further enhanced the capabilities of text-to-image synthesis.

Removal for Generated Images

Image removal techniques focus on eliminating unwanted elements like specific objects, logos, or watermarks from digital images. Recent advancements, particularly in deep learning and generative models, have improved these processes, ensuring better image quality and realism.

Concept Removal Concept removal is essential for privacy preservation, content moderation, and augmented reality. Recent frameworks using adversarial training and generative modeling effectively suppress sensitive information while maintaining image fidelity (Gandikota et al. 2024; Wang et al. 2023; Tsai et al. 2023; Hong, Lee, and Woo 2024). Advanced methods combine semantic segmentation with generative models to obfuscate specific objects (Pham et al. 2023, 2024; Li et al. 2024; Zhao et al. 2024; Xiong et al. 2024), and attention-based approaches enhance privacy by dynamically suppressing salient regions (Yang, Mu, and Deng 2022).

Watermark Removal Watermark removal is vital for repurposing images and videos legally. Traditional signal processing methods like frequency filtering often degraded image quality (Ray and Roy 2020). However, deep learning techniques, including CNNs and autoencoders, now enable more effective watermark removal and content reconstruction (Chen et al. 2021). Despite these advances, ethical and legal concerns remain, driving the development of methods that comply with copyright laws (Singh, Jain, and Sharma 2013). Recent techniques focus on improving accuracy and minimizing artifacts through adversarial training and multi-scale analysis (Luo et al. 2023).

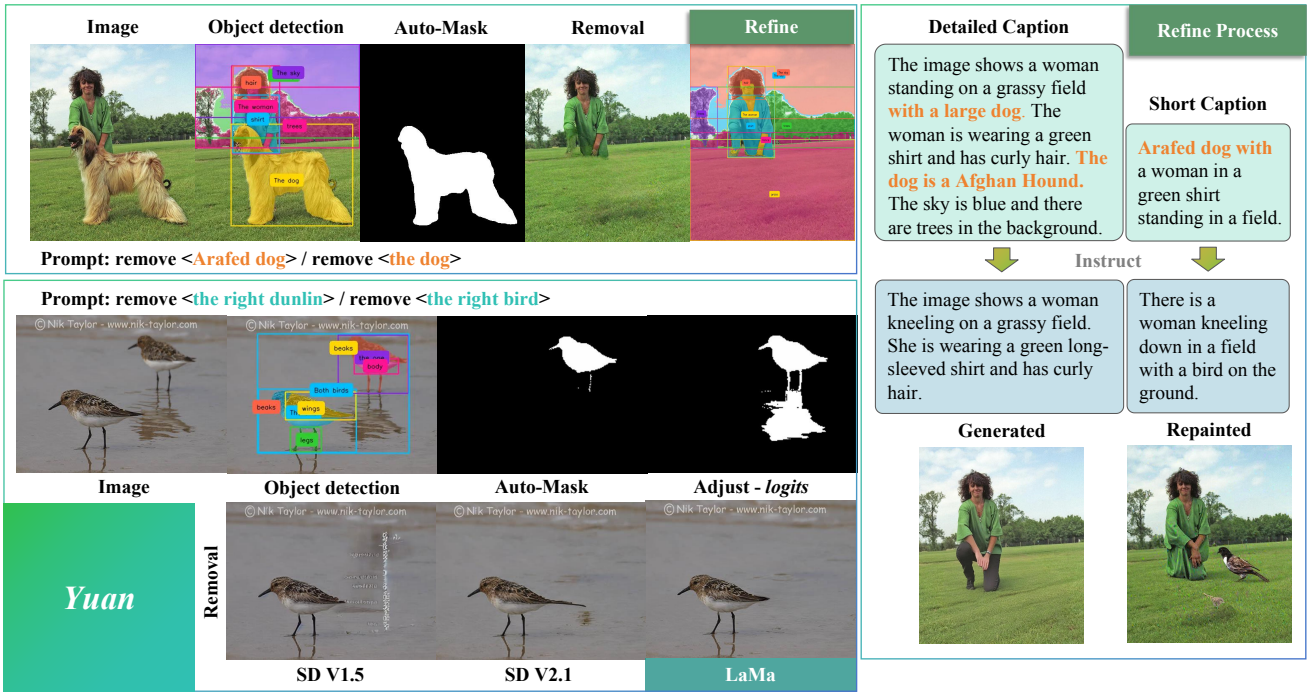

Figure 2: Our *Yuan* framework: (a) Object detection by user prompt, (b) Automatic mask generation, (c) Object removal, (d) Inpainting and preserving original context, and (e) Refined image.

Ongoing developments in concept and watermark removal are expanding the possibilities in digital image processing, with future efforts aimed at increasing robustness and adaptability across various applications.

Image Editing and Inpainting

Conditional image editing and inpainting have advanced significantly, enabling applications in content creation, image restoration, and augmented reality. However, challenges remain, particularly in handling shadows, complex scenes, and enhancing user interaction.

Shadow Handling Shadows pose difficulties due to their complex interplay with light sources, objects, and backgrounds. Many algorithms struggle with light and shadow consistency, often leading to distortions when shadows are improperly removed (Le and Samaras 2019). Additionally, shadows typically have gradient edges, but current methods often produce unnatural hard edges or artifacts during editing (Liu et al. 2021).

Editing Complex Scenes Editing complex scenes, especially those with multiple objects, requires algorithms to maintain spatial relationships and scene coherence. Current techniques often fail to preserve local and global consistency, resulting in edited areas that clash with the original image's color, texture, or lighting (Wang et al. 2020; Zhang and Schomaker 2021).

User Interaction User interaction in conditional editing tools is still limited. Systems often struggle to understand user intent, leading to results that do not align with expectations (Borch and Hee Min 2022). Additionally, many systems lack real-time feedback and dynamic adjustment capabilities, requiring users to make cumbersome manual adjustments (Sun et al. 2022).

Our Work

Distinct from existing methodologies, our *Yuan* framework introduces a revolutionary automated approach to identify and correct visual imperfections by seamlessly integrating text and image data. By advancing beyond the traditional reliance on manual masking, our framework employs advanced segmentation and inpainting modules, significantly enhancing the efficiency and effectiveness of the image refinement process. This automation not only aligns with the latest developments in generative AI, but also addresses critical gaps identified in current practices, such as the labor-intensive nature of manual interventions and the inconsistencies they introduce.

Proposed Method - *Yuan*

Overview

As illustrated in Fig. 2, given a synthetic image generated by any text-to-image (T2I) model, our proposed method, *Yuan*, employs the models to conditionally analyze the prompt and automatically generate segmentation masks. These masks

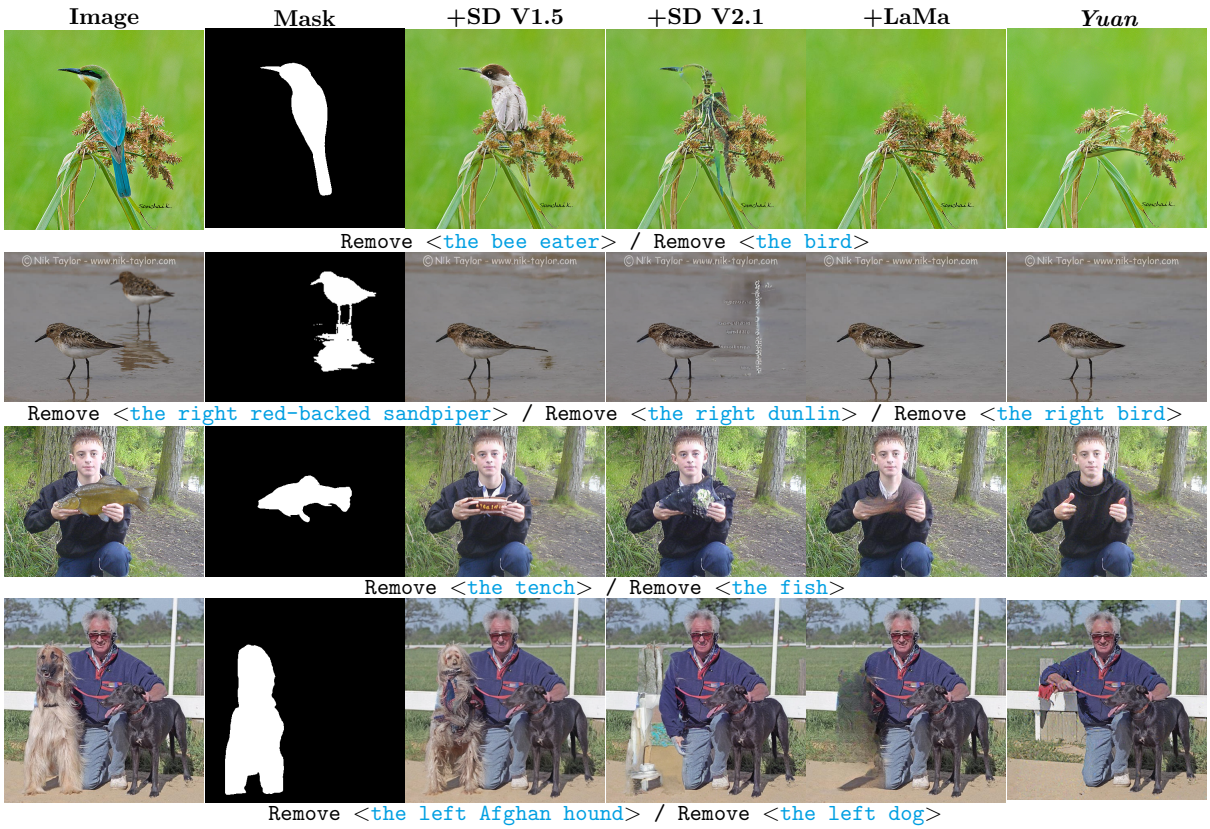

Figure 3: A comparison among Grounded SAM+SD V1.5, +SD V2.1, +LaMa, and *Yuan* for different text prompt.

Metrics	ImageNet100				Stanford-dogs				Generated-cats			
	Image	+SD	+LaMa	*Yuan*	Image	+SD	+LaMa	*Yuan*	Image	+SD	+LaMa	*Yuan*
NIQE↓	*3.7425*	5.2829	3.0905	**3.0890**	*3.3380*	4.6187	4.0785	**3.4691**	*5.2829*	6.2217	**5.0716**	5.2465
BRISQUE↓	*26.6525*	32.1852	**24.7853**	25.6086	*9.6086*	25.1237	22.0275	**16.2062**	*32.1852*	**37.4372**	45.6096	39.4333
PI↓	*2.5558*	5.9084	2.0921	**2.0124**	*2.2204*	3.2685	2.6190	**2.2841**	*5.9084*	6.7089	5.5097	**5.4679**

Table 1: Comparison of object removal performance across different models. It compares the performance of Grounded SAM+SD, +LaMa, and *Yuan* on object removal tasks across three datasets: ImageNet100, Stanford-dogs, and Generated-cats.

are then used to selectively preserve or modify specific regions of the image, ensuring the integrity and coherence of the original visual content. The process integrates the strengths of advanced segmentation and object detection models, followed by a robust inpainting approach to maintain the original context as illustrated in Algorithm 1.

Automatic Mask Generation

In order to create an automated process for generating masks based on the synthesis prompt, *Yuan* utilised grounded SAM. This is because it combines the precise object detection capabilities of Grounding DINO with the powerful segmentation abilities of the Segment Anything Model (SAM). This integration removes the need for manual intervention. The process begins with Grounding DINO, which uses a transformer-based architecture to detect objects in detail. Its loss function, ($\mathcal{L}_{GDINO}$), includes components for both classification ($\mathcal{L}_{cls}$) and localization ($\mathcal{L}_{loc}$), ensuring accurate detection. Once the objects are detected, the SAM model takes over to segment the identified regions. By conditioning on both the synthesis prompt and image features, Grounded SAM will auto generate precise segmentation masks (M_{SAM}). This automated approach addresses the limitations of manual methods, enhancing consistency and precision in identifying regions of interest (details see Appendix).

Inpainting for Image Preservation

To preserve the original characteristics of the image, we adopt the LaMa Inpainting model over traditional diffusion-based methods. This is because diffusion-based techniques often introduce inconsistencies and artifacts that can detract from the image's coherence. In contrast, the LaMa model focuses on inpainting, which involves restoring specific masked regions based on the surrounding context. The LaMa model is optimized to inpaint large masked regions ef-

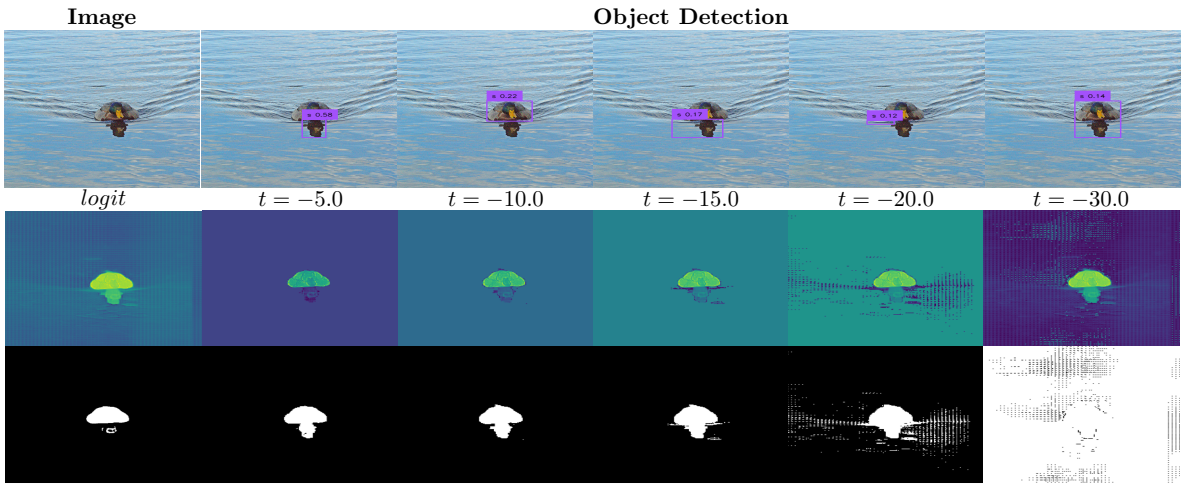

Figure 4: Ablation study results on $logits$ threshold (t) adjustment for automatic mask generation.

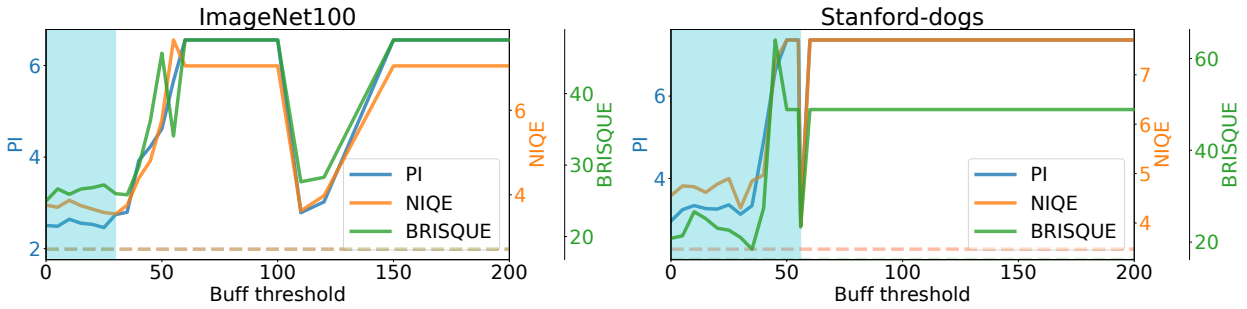

Figure 5: Ablation study results of threshold in buffer (b) and $logits$ (t). For different datasets, the threshold needs to be adjusted as needed. For *Yuan*, we recommend adding two adjustable parameters, exposure b and t, based on the original settings. This will provide convenience and service for different generated images to achieve the best results. Full figures see Fig. A3.

fectively, predicting and filling these areas while maintaining visual consistency. This process is governed by a loss function ($\mathcal{L}_{inpaint}$) that balances reconstruction and perceptual similarity. $\mathcal{L}_{recon}$ ensures the inpainted region matches the original image's appearance, and $\mathcal{L}_{perc}$ maintains high-level perceptual similarity. The parameter β controls the balance between these two objectives. This inpainting approach ensures that modified regions blend seamlessly with untouched areas, maintaining the original image's visual integrity and coherence (detail in Appendix).

$$I_{inpaint} = \text{LaMa}(I, M_{SAM}) \tag{1}$$

Refining Visual Imperfects

Our approach to refining visual imperfections consists of two key steps: (i) adjusting the output logits of the SAM to obtain more accurate masks, and (ii) employing Prompt-to-Prompt techniques for image repainting.

Adjusting Logits for Improved Masks Mask generation is vital for identifying regions of interest in image processing. Initially, a threshold $t = 0$ was used, but it often missed out shadowed areas, leading to incomplete masks. Lowering the threshold t improved feature coverage:

$$\Delta M_{SAM}(x) = \begin{cases} 1, & \text{if } logit(x) \geq t \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Experiments demonstrated that setting $t = -10$ typically produces the best results, although the optimal value may vary depending on image complexity. This threshold is adjustable by the user, allowing for improved mask coverage, especially in images with complex backgrounds.

Repainting via Prompt Instruct When the adjusted masks fall short of the desired refinement, we employ Prompt-to-Prompt techniques for further optimization, guiding the model by analyzing semantic differences between the original and repainted images. To create a high-quality training dataset, we use Florence2 (Xiao et al. 2024) to generate captions that provide detailed semantic information, enabling the identification of differences between the original (C_I) and repainted (C_r) images. This data, combined with user modification requests (P), is used to fine-tune a GPT model to map between the original and inpainted captions, resulting in a refined caption (ΔC_r) that serves as the prompt for T2I models such as Stable Diffusion.

Algorithm 1: *Yuan* - Object Removal

Require: Synthetic image I from any T2I model
Prompt P from user input
Ensure: Refined image $output$
$D_{GDINO} \leftarrow$ GDINO(I, P) {Detect objects}
$M_{SAM} \leftarrow$ SAM(D_{GDINO}) {Generate mask}
$I_{masked} \leftarrow$ Apply ΔM_{SAM} to I
$I_{inpaint} \leftarrow$ LaMa$(I_{masked}, \Delta M_{SAM})$ {Inpaint}
$output \leftarrow I_{inpaint}$
if $I_{inpaint}$ is insufficient **then**
  $\Delta M_{SAM} \leftarrow logit(t)$ {Adjust mask}
  $\Delta I_{masked} \leftarrow$ Apply ΔM_{SAM} to I
  $I_{inpaint2} \leftarrow$ LaMa$(\Delta I_{masked}, \Delta M_{SAM})$ {Inpaint}
  $output \leftarrow I_{inpaint2}$
  if $I_{inpaint2}$ is insufficient **then**
    $C_I \leftarrow$ Caption(I) {Generate caption}
    $C_r \leftarrow \text{GPT}_{\text{fine-tuned}}(P, C_I)$ {Generate new caption}
    $I_{refined} \leftarrow$ Generate$(\Delta C_r, I)$
    $output \leftarrow I_{refined}$
  end if
end if
return $output$

$$\text{Caption}(I) \rightarrow C_I, \quad \text{Caption}(I_{inpaint}) \rightarrow C_r \tag{3}$$

$$GPT_{ft}(P, C_I) \rightarrow \Delta C_r \tag{4}$$

Finally, the fine-tuned GPT_{ft} interprets and executes image optimization instructions by generating captions that guide Stable Diffusion in producing refined images. Empirically, we found that this approach improves visual consistency and quality, especially in cases that require extensive content modification, ensuring that the final images align with the user's intent and aesthetic objectives.

Experiment

Dataset Description

We conducted experiments using three datasets: ImageNet100 (Russakovsky et al. 2015), Stanford-dogs (Khosla et al. 2011), and Generated-cats. ImageNet100, a subset of ImageNet1K, includes 100 categories with 60,000 training images and 10,000 validation images, providing a condensed but representative dataset for model evaluation. Stanford-dogs contains 120 dog breeds with 20,580 images, designed for fine-grained classification. Generated-cats is a dataset created using a stable diffusion model with the prompt $<$ `cat` $>$ (details in Appendix).

Experimental Settings

The experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. For object detection, Grounding DINO used initial box and text thresholds of 0.1. In the ImageNet100 dataset, classification labels were used as text prompts, while in the Stanford-dogs dataset, either classification labels or the keyword `<dog>` were used for object removal. To evaluate generalization, 100 "cat"-related sentences were generated using ChatGPT-3.5 (details in the Appendix) and used to create a custom dataset of 100 images via Stable Diffusion, with the generation process taking about 0.5 hours. The GPT model used for fine-tuning was gpt-4o-mini.

Image size	BLIP	Florence2	SD T2I	SD I2I	SD inpaint
512×512	27.53	6.04	7.99	12.55	10.92

Table 2: Inference time. Unit: second/image.

Evaluation Metrics

We use three no-reference image quality assessment metrics to evaluate perceptual image quality. **NIQE** assesses image naturalness and distortion based on natural scene statistics, with lower scores indicating better quality (Mittal, Soundararajan, and Bovik 2012). **BRISQUE** analyzes spatial domain features to quantify distortion, where lower scores also indicate higher quality (Mittal, Soundararajan, and Bovik 2012). **PI** combines NIQE and BRISQUE scores to provide an overall quality measure, with lower scores reflecting better perceptual quality (Wang et al. 2004). These metrics allow for comprehensive evaluation across different distortion scenarios.

Comparison

Quantitative Analysis Table 1 compares three models (Grounded-SAM+SD, Grounded-SAM+LaMa, and *Yuan*) across the ImageNet100, Stanford-dogs, and Generated-cats datasets, focusing on removal quality using NIQE, BRISQUE, and PI metrics. *Yuan* consistently outperforms the other models, particularly on the Stanford-dogs dataset, with LaMa following and SD performing the worst. NIQE and PI scores show that LaMa and *Yuan* produce images with naturalness and realism close to the original. In BRISQUE, LaMa and *Yuan* surpass the original images on ImageNet100, indicating high-quality outputs with fewer distortions. However, performance declines on the other datasets, especially Generated-cats, likely due to inherent distortions. Despite these challenges, *Yuan* remains closest to the original images, demonstrating robustness and adaptability across different datasets.

Inference Time We compared the inference times of the models used in this study (Table 2). Florence2 is approximately 4.56 times faster than BLIP for image caption generation. Among the text-to-image (T2I) models, SD is the fastest, but its results are suboptimal. Our *Yuan* framework balances operational efficiency and system overhead, leading us to choose a combination of Florence2 and SD inpaint.

Qualitative Analysis Figs. 3 and A.2 compare the performance of SD V1.5, SD V2.1, LaMa, and *Yuan* across various text prompts for object removal. For the prompt `Remove the <dog>`, SD models leave artifacts and incomplete blending, while *Yuan* effectively removes the object with minimal artifacts, preserving texture. In more complex scenes, SD models struggle with context, and LaMa

Figure 6: Limitations of *Yuan*. The challenge of accurately rendering human hands due to complex anatomy, and the generation of unintended content during the refinement process.

performs slightly better, but *Yuan* excels in maintaining context without distortion. For the prompt `Remove the <hand>`, SD models leave visible traces, and while LaMa improves on this, *Yuan* successfully removes the object, ensuring natural appearance. Similarly, for `Remove the <trunk>`, SD models produce artifacts, and LaMa lacks fine detail handling, but *Yuan* achieves clean removal and preserves texture. Overall, *Yuan* consistently outperforms the other models, accurately removing objects with minimal artifacts, demonstrating significant advancements in text-guided image editing.

Ablation Study

Buffer Zone The buffer zone acts as a transitional area around the removal region, smoothing edges and reducing artifacts to improve the reconstructed image's quality and natural appearance. Table A.1 and Figs. 5 and A.1 show the impact of varying buffer thresholds (b) from 0 to 200 across different datasets. Optimal ranges are as follows: ImageNet100 (0∼30), Stanford-dogs (0∼56), and Generated-cats (50∼60). Results indicate diminishing returns for $b > 50$, as metrics like NIQE, BRISQUE, and PI stabilize. For Generated-cats, a higher b improves clarity due to feathering effects and ambiguous boundaries. While increasing b generally enhances quality, it may also reduce original detail. The study identifies optimal thresholds to balance reconstruction quality and content originality, with $b = 15$ set as the default in uncertain cases.

Binarize Adjuster The logit threshold (t) is crucial for determining segmentation mask sensitivity, impacting the precision of object removal. Optimal t values are as follows: ImageNet100 ($-6 \sim -5$), Stanford-dogs (-8), and Generated-cats ($-14 \sim -7$). Figs. 4 and 5, and Table A.2 show that the best performance on NIQE, BRISQUE, and PI metrics occurs when $-10 \leq t \leq 0$. Below $t = -10$, quality degrades significantly, with sharp deterioration at $t \leq -15$. ImageNet100 and Stanford-dogs datasets show stable metric changes, while Generated-cats exhibit more variability due to the quality of generated images. Fine-tuning t is essential for balancing image quality and reconstruction effectiveness, with $t = -10$ set as the default for automation. This adjuster is key to enhancing naturalness and realism while minimizing artifacts.

Limitations

Despite promising results, *Yuan* has two limitations (Fig. 6): **Hand generation in generated models:** Current generated models struggle with rendering human hands accurately due to their complex anatomy and variable poses, often leading to artifacts. Enhancing hand generation fidelity remains a significant challenge, requiring improvements in both model architecture and training data. **Unintended content generation after refinement:** While effective at object removal and refinement, the process can sometimes introduce unintended elements, requiring additional refinement rounds. It can be resource-intensive and impact efficiency, highlighting the need for better controls during generation to prevent such occurrences. Addressing these limitations is essential for improving the robustness and reliability of our framework. Future work should focus on enhancing generated capabilities, particularly in generating complex anatomical features, and refining processes to better meet user expectations.

Conclusion

Text-to-image synthesis has made significant strides, but generated images often suffer from visual imperfections like anatomical inconsistencies and unwanted textual elements. Traditional correction methods relying on manual masks are time-consuming and inconsistent. This paper introduces *Yuan*, a framework that automatically addresses these visual flaws by integrating a grounded segmentation module and an inpainting module. *Yuan* effectively identifies and corrects image imperfections without the need for manual intervention, ensuring visual and contextual coherence. Extensive evaluations demonstrate *Yuan*'s robustness and effectiveness, making it a valuable contribution to enhancing the quality and practicality of text-to-image synthesis.

Appendix

Appendix of this paper can be found at `https://github.com/YuZhenyuLindy/Yuan.git`

References

Borch, C.; and Hee Min, B. 2022. Toward a sociology of machine learning explainability: Human–machine interaction in deep neural network-based automated trading. *Big Data & Society*, 9(2): 20539517221111361.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Cetinic, E.; and She, J. 2022. Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2): 1–22.

Chen, X.; Wang, W.; Bender, C.; Ding, Y.; Jia, R.; Li, B.; and Song, D. 2021. Refit: a unified watermark removal framework for deep learning systems with limited data. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, 321–335.

Gafni, O.; Polyak, A.; Ashual, O.; Sheynin, S.; Parikh, D.; and Taigman, Y. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, 89–106. Springer.

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*.

Gandikota, R.; Orgad, H.; Belinkov, Y.; Materzyńska, J.; and Bau, D. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5111–5120.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

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

Hong, S.; Lee, J.; and Woo, S. S. 2024. All but One: Surgical Concept Erasing with Model Preservation in Text-to-Image Diffusion Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21143–21151.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Li, F.-F. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPR-W*, volume 2. Citeseer.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Le, H.; and Samaras, D. 2019. Shadow removal via shadow image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8578–8587.

Li, S.; van de Weijer, J.; Hu, T.; Khan, F. S.; Hou, Q.; Wang, Y.; and Yang, J. 2024. Get What You Want, Not What You Don't: Image Content Suppression for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2402.05375*.

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 *CVPR*, 22511–22521.

Liu, Z.; Yin, H.; Wu, X.; Wu, Z.; Mi, Y.; and Wang, S. 2021. From shadow generation to shadow removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4927–4936.

Luo, X.; Li, Y.; Chang, H.; Liu, C.; Milanfar, P.; and Yang, F. 2023. DVMark: a deep multiscale framework for video watermarking. *IEEE Transactions on Image Processing*.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3): 209–212.

Pham, M.; Marshall, K. O.; Cohen, N.; Mittal, G.; and Hegde, C. 2023. Circumventing concept erasure methods for text-to-image generative models. In *The Twelfth International Conference on Learning Representations*.

Pham, M.; Marshall, K. O.; Hegde, C.; and Cohen, N. 2024. Robust Concept Erasure Using Task Vectors. *arXiv preprint arXiv:2404.03631*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Ray, A.; and Roy, S. 2020. Recent trends in image watermarking techniques for copyright protection: a survey. *International Journal of Multimedia Information Retrieval*, 9(4): 249–270.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Singh, N.; Jain, M.; and Sharma, S. 2013. A survey of digital watermarking techniques. *International Journal of Modern Communication Technologies and Research*, 1(6): 265852.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Sun, J.; Wang, X.; Shi, Y.; Wang, L.; Wang, J.; and Liu, Y. 2022. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Transactions on Graphics (ToG)*, 41(6): 1–10.

Tsai, Y.-L.; Hsu, C.-Y.; Xie, C.; Lin, C.-H.; Chen, J.-Y.; Li, B.; Chen, P.-Y.; Yu, C.-M.; and Huang, C.-Y. 2023. Ring-A-Bell! How Reliable are Concept Removal Methods for Diffusion Models? *arXiv preprint arXiv:2310.10012*.

Wang, P.; Yang, Y.; and Yu, Z. 2024. Multi-batch Nuclear-norm Adversarial Network for Unsupervised Domain Adaptation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.

Wang, T.; Zhang, Y.; Qi, S.; Zhao, R.; Xia, Z.; and Weng, J. 2023. Security and privacy on generative data in aigc: A survey. *arXiv preprint arXiv:2309.09435*.

Wang, Y.; Huang, W.; Sun, F.; Xu, T.; Rong, Y.; and Huang, J. 2020. Deep multimodal fusion by channel exchanging. *Advances in neural information processing systems*, 33: 4835–4845.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wu, J.; Gan, W.; Chen, Z.; Wan, S.; and Lin, H. 2023. Ai-generated content (aigc): A survey. *arXiv preprint arXiv:2304.06632*.

Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; and Yuan, L. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4818–4829.

Xiong, T.; Wu, Y.; Xie, E.; Li, Z.; and Liu, X. 2024. Editing Massive Concepts in Text-to-Image Diffusion Models. *arXiv preprint arXiv:2403.13807*.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1316–1324.

Yang, Y.; Mu, K.; and Deng, R. H. 2022. Lightweight privacy-preserving GAN framework for model training and image synthesis. *IEEE Transactions on Information Forensics and Security*, 17: 1083–1098.

Yao, B. Z.; Yang, X.; Lin, L.; Lee, M. W.; and Zhu, S.-C. 2010. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8): 1485–1508.

Yu, Z. 2024. Improved implicit diffusion model with knowledge distillation to estimate the spatial distribution density of carbon stock in remote sensing imagery. *arXiv preprint arXiv:2411.17973*.

Yu, Z.; Wang, J.; Yang, X.; and Ma, J. 2023. Superpixel-Based Style Transfer Method for Single-Temporal Remote Sensing Image Identification in Forest Type Groups. *Remote Sensing*, 15(15): 3875.

Yu, Z.; and Wang, P. 2024. CaPAN: Class-aware Prototypical Adversarial Networks for Unsupervised Domain Adaptation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.

Zhang, Z.; and Schomaker, L. 2021. Dtgan: Dual attention generative adversarial networks for text-to-image generation. In *2021 international joint conference on neural networks (IJCNN)*, 1–8. IEEE.

Zhao, M.; Zhang, L.; Zheng, T.; Kong, Y.; and Yin, B. 2024. Separable Multi-Concept Erasure from Diffusion Models. *arXiv preprint arXiv:2402.05947*.