Automated Virtual Product Placement and Assessment in Images using Diffusion Models

Mohammad Mahmudul Alam*
University of Maryland, Baltimore County
Baltimore, MD, USA
m256@umbc.edu

Negin Sokhandan
Amazon Web Services (AWS)
Santa Clara, CA, USA
ngnsl@amazon.com

Emmett Goodman
Amazon Web Services (AWS)
Santa Clara, CA, USA
edmgood@amazon.com

Abstract

In Virtual Product Placement (VPP) applications, the discrete integration of specific brand products into images or videos has emerged as a challenging yet important task. This paper introduces a novel three-stage fully automated VPP system. In the first stage, a language-guided image segmentation model identifies optimal regions within images for product inpainting. In the second stage, Stable Diffusion (SD), fine-tuned with a few example product images, is used to inpaint the product into the previously identified candidate regions. The final stage introduces an 'Alignment Module', which is designed to effectively sieve out low-quality images. Comprehensive experiments demonstrate that the Alignment Module ensures the presence of the intended product in every generated image and enhances the average quality of images by 35%. The results presented in this paper demonstrate the effectiveness of the proposed VPP system, which holds significant potential for transforming the landscape of virtual advertising and marketing strategies.

1. Introduction

Virtual Product Placement (VPP) refers to the unobtrusive, digital integration of branded products into visual content, which is often employed as a stealth marketing strategy [15]. Advertising solutions utilizing VPP have significant appeal due to their high customizability, effectiveness across diverse customer bases, and quantifiable efficiency.

(a) Background (b) Inpainting

Figure 1. An illustration of the proposed VPP system with an Amazon Echo Dot device. The input background image is shown in (a), and the inpainted output image is shown in (b) where an Amazon Echo Dot device is placed on the kitchen countertop by automatic identification of optimal location.

Previous research underscores the impact of product placement within realms such as virtual reality [22] and video games [5]. With the recent advancements in generative AI technologies, the potential for product placement has been further expanded through the utilization of diffusion models. Significant research has focused on the development of controlled inpainting via diffusion models, albeit largely without an explicit emphasis on advertising applications [1, 8, 11]. However, these methods can be fine-tuned with a small set of 4 to 5 product sample images to generate high-quality advertising visual content.

In this paper, we propose a novel, three-stage, fully automated system that carries out semantic inpainting of products by fine-tuning a pre-trained Stable Diffusion (SD) model [18]. In the first stage, a suitable location is identified for product placement using visual question answering and text-conditioned instant segmentation. The output of this stage is a binary mask highlighting the identified location. Subsequently, this masked region undergoes inpainting using a fine-tuned SD model. This SD model is fine-tuned by

*The author performed this work as an intern at Amazon Web Services (AWS). Accepted at 6th AI for Content Creation (AI4CC) workshop at CVPR 2024. (Preprint)

DreamBooth [19] approach utilizing a few sample images of the product along with a unique identifier text prompt. Finally, the quality of the inpainted image is evaluated by a proposed Alignment Module, a discriminative method that measures the image quality, or the alignment of the generated image with human expectations. An illustration of the proposed VPP system is presented in Figure 1 with an Amazon Echo Dot device.

Controlled inpainting of a specific product is a challenging task. For example, the model may fail to inpaint the intended object at all. If a product is indeed introduced through inpainting, the product created may not be realistic and may display distortions of shape, size, or color. Similarly, the background surrounding the inpainted product may be altered in such a way that it either meaningfully obscures key background elements or even completely changes the background image. This becomes especially problematic when the background images contain human elements, as models can transform them into disturbing visuals. As a result, the proposed Alignment Module is designed to address these complications, with its primary focus being on the appearance, quality, and size of the generated product.

To exert control over the size of the generated product, morphological transformations, specifically erosion, and dilation, are employed. By adjusting the size of the mask through dilation or erosion, the size of the inpainted product can be effectively increased or decreased. This allows the system to generate a product of an appropriate size.

In summary, the main contributions of this paper are twofold. The first pertains to the design of a fully automated Virtual Product Placement (VPP) system capable of generating high-resolution, customer-quality visual content. The second involves the development of a discriminative method that automatically eliminates subpar images, premised on the content, quality, and size of the product generated.

The remainder of this paper is organized as follows. In section 2 we will delve into the related literature, with a specific emphasis on semantic inpainting methods utilizing diffusion models, and section 3 will highlight the broad contributions of the paper. Next, the proposed end-to-end pipeline for automatic VPP will be discussed in section 4. This includes a detailed examination of the three primary stages of the solution, along with the three sub-modules of the Alignment Module. Thereafter, we will elucidate the experimental design and evaluation methodologies adopted and report the corresponding results in section 5. Subsequently, deployment strategy and web application design will be explained in section 6. Finally, the paper will conclude with an outline of the identified limitations of our proposed methodology in section 7, complemented by a discussion on potential avenues for future research.

2. Related Works

Recently, there has been significant progress in developing semantic or localized image editing using diffusion models - largely without an explicit focus on digital marketing. Nevertheless, new generative AI approaches promise significant advances in VPP technology. For instance, in Blended Diffusion [1], the authors proposed a method of localized image editing using image masking and natural language. The area of interest is first masked and then modified using a text prompt. The authors employed a pre-trained CLIP model [17] along with pre-trained Denoising Diffusion Probabilistic Models (DDPM) [7] to generate natural images in the area of interest.

Similar to Blended Diffusion, Couairon et. al. [3] proposed a method of semantic editing with a mask using a diffusion model. However, instead of taking the mask from the user, the mask is generated automatically. Nevertheless, a text query input from the user is utilized to generate the mask. The difference in noise estimates, as determined by the diffusion model based on the reference text and the query text, is calculated. This difference is then used to infer the mask. The image is noised iteratively during the forward process and in the reverse Denoising Diffusion Implicit Model (DDIM) [21] steps, the denoised image is interpolated with the same step output of the forward process using masking.

Paint by Word proposed by Bau et. al. [2], is also similar, however instead of a diffusion model they utilized a Generative Adversarial Networks (GAN) [4] with a mask for semantic editing guided by text. On the other hand, Imagic [8] also performs text-based semantic editing on images using a diffusion model but without using any mask. Their approach consists of three steps. In the beginning, text embedding for a given image is optimized. Then the generative diffusion model is optimized for the given image with fixed-optimized text embedding. Finally, the target and optimized embedding are linearly interpolated to achieve input image and target text alignment. Likewise, a semantic editing method using a pre-trained text-conditioned diffusion model focusing on the mixing of two concepts is proposed by [12]. In this method, a given image is noised for several steps and then denoised with text condition. During the denoising process, the output of a denoising stage is also linearly interpolated with the output of a forward noise mixing stage.

Hertz et. al. [6] took a different approach to semantic image editing where text and image embeddings are fused using cross-attention. The cross-attention maps are incorporated with the Imagen diffusion model [20]. However, instead of editing any given image, their approach edits a generated image using a text prompt which lacks any interest when VPP is concerned. Alternatively, Stochastic Differential Edit (SDEdit) [16] synthesizes images from stroke

paintings and can edit images based on stroke images. For image synthesis, coarse colored strokes are used and for editing, colored stroke on real images or image patches on target images is used as a guide. It adds Gaussian noise to an image guide of a specific standard deviation and then solves the corresponding Stochastic Differential Equations (SDE) to produce the synthetic or edited image.

To generate images from a prompt in a controlled fashion and to gain more control over the generated image, Li et. al proposed grounded text-to-image generation (GLIGEN) [11]. It feeds the model the embedding of the guiding elements such as bounding boxes, key points, or semantic maps. Using the same guiding components, inpainting can be performed in a target image.

DreamBooth [19] fine-tunes the pre-trained diffusion model to expand the dictionary of the model for a specific subject. Given a few examples of the subject, a diffusion model such as Imagen [20] is fine-tuned using random samples generated by the model itself and new subject images by optimizing a reconstruction loss. The new subject images are conditioned using a text prompt with a unique identifier. Fine-tuning a pre-trained diffusion model with a new subject is of great importance in the context of VPP. Therefore, in this paper DreamBooth approach is utilized to expand the model's dictionary by learning from a few sample images of the product.

3. Contributions

In this paper, a method of automated virtual product placement and assessment in images using diffusion models is designed. Our broad contributions are as follows:

1. We introduce a novel fully automated VPP system that carries out automatic semantic inpainting of the product in the optimal location using language-guided segmentation and fine-tuned stable diffusion models.
2. We proposed a cascaded three-stage assessment module named 'Alignment Module' designed to sieve out low-quality images that ensure the presence of the intended product in every generated output image.
3. Morphological transformations are employed such as dilation and erosion to adjust the size of the mask, therefore, to increase or decrease the size of the inpainted product allowing generating a product of appropriate size.
4. Experiments are performed to validate the results by blind evaluation of the generated images with and without the Alignment module resulting in 35% improvement in average quality.
5. The inpainted product generated by the proposed system is not only qualitatively more realistic compared to the previous inpainting approach [23] but also shows a superior quantitative CLIP score.

4. Methodology

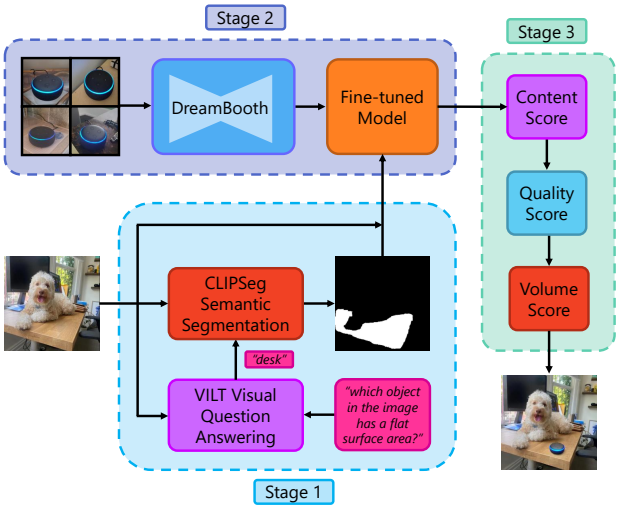


Figure 2. The block diagram of the proposed solution for the VPP system where each of the three stages is distinguished by varied color blocks. In stage 1, a suitable placement for product inpainting is determined by creating a mask using CLIPSeg and VILT models. Next, in stage 2, semantic inpainting is performed in the masked area using the fine-tuned DreamBooth model. Finally, stage 3 contains the cascaded sub-modules of the Alignment Module to discard low-quality images.

4.1. Proposed Method

For semantic inpainting, we utilized the DreamBooth algorithm [19] to fine-tune stable diffusion using five representative images of the product and a text prompt with a unique identifier. Even with a limited set of five sample images, the fine-tuned DreamBooth model was capable of generating images of the product integrated with its background.

Nevertheless, when inpainting was conducted with this fine-tuned model, the resulting quality of the inpainted product was significantly compromised. To enhance the quality of the product in the inpainted image, we augmented the sample images through random scaling and random cropping, consequently generating a total of 1,000 product images used to fine-tune SD.

4.2. Product Localization Module

The proposed VPP system operates in three stages. A core challenge in product placement lies in pinpointing a suitable location for the item within the background. In the first stage, this placement is indicated via the generation of a binary mask. To automate this masking process, we leveraged the capabilities of the Vision and Language Transformer (ViLT) Visual Question Answering (VQA) model [9] in conjunction with the Contrastive Language-

Image Pretraining (CLIP) [17]-based semantic segmentation method, named CLIPSeg [13]. Notably, each product tends to have a prototypical location for its placement. For example, an optimal location for an Amazon Echo Dot device is atop a flat surface, such as a desk or table. Thus, by posing a straightforward query to the VQA model, such as "Which object in the image has a flat surface area?", we can pinpoint an appropriate location for the product. Subsequently, the identified location's name is provided to the CLIPSeg model, along with the input image, resulting in the generation of a binary mask for the object.

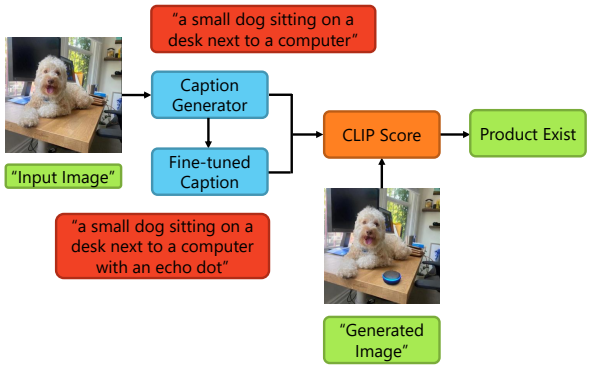
4.3. Product Inpainting Module

In the second stage, the input image and the generated binary mask are fed to the fine-tuned DreamBooth model to perform inpainting on the masked region. Product inpainting presents several challenges: the product might not manifest in the inpainted region; if it does, its quality could be compromised or distorted, and its size might be disproportionate to the surrounding context. To systematically detect these issues, we introduce the third stage: the Alignment Module.

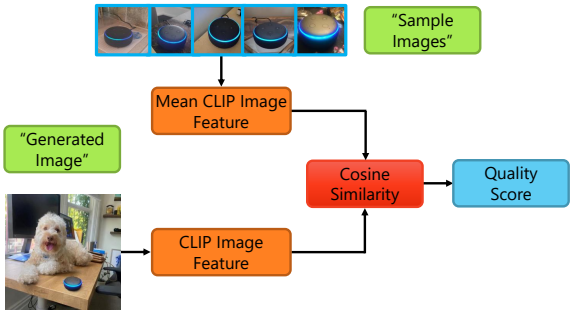
4.4. Product Alignment Module

The Alignment Module comprises three sub-modules: Content, Quality, and Volume. The Content sub-module serves as a binary classifier, determining the presence of the product in the generated image. If the product's probability of existence surpasses a predefined threshold, then the Quality score is calculated for that image. This score evaluates the quality of the inpainted product in relation to the sample images originally used to train the SD model. Finally, if the image's quality score exceeds the set quality threshold, the Volume sub-module assesses the product's size in proportion to the background image. The generated image will be successfully accepted and presented to the user only if all three scores within the Product Quality Alignment Module meet their respective thresholds.

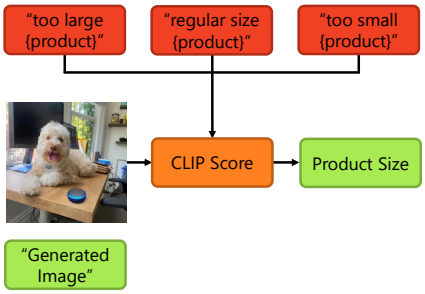
Within the Content module, an image captioning model [14] is employed to generate a caption, which is then refined by incorporating the product's name. The super-class name of the product can also be utilized. Both the captions and the inpainted image are fed into the CLIP model to derive a CLIP score. If the modified caption scores above 70%, it's inferred that the product exists in the inpainted image. The Quality module contrasts the mean CLIP image features of the sample images with the CLIP image feature of the generated image. The greater the resemblance of the inpainted product to the sample images, the higher the quality score. A threshold of 70% has been established. The Volume module finally gauges the size of the inpainted product. The generated image is processed through the CLIP model, accompanied by three distinct textual size prompts. Given that



(a) Content Sub-module



(b) Quality Sub-module



(c) Volume Sub-module

Figure 3. Block diagram of each of the components of the Alignment Module. The Content sub-module is built using a pre-trained caption generator and CLIP models shown in (a). The generated caption is fine-tuned by adding the name of the intended product to the caption. For the Quality sub-module, the image features of the same CLIP model are utilized shown in (b). Finally, in the Volume sub-module, the same CLIP model with three different size text prompts is used shown in (c).

size perception can be subjective and varies based on camera proximity, a milder threshold of 34% (slightly above a random guess) has been selected. The comprehensive block diagram of the proposed VPP system is illustrated in Figure 2, with the three stages distinguished by varied color blocks. The block diagrams for each sub-module can be found in Figure 3.

The Volume sub-module provides insights regarding the size of the inpainted product. To modify the product's size, the mask's dimensions must be adjusted. For this task, morphological transformations, including mask erosion and dilation, can be employed on the binary mask. These transformations can either reduce or augment the mask area, allowing the inpainting module to produce a product image of the desired size. The relationship between alterations in the mask area and the size of the inpainted product across various erosion iterations is depicted in Figure 4. Approximately, 25 iterations of erosion consume around 3 milliseconds, making it highly cost-effective.

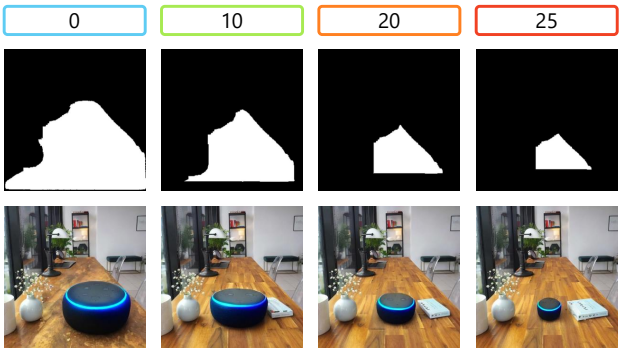


Figure 4. Application of erosion to the mask where a kernel of size (5×5) is used for 0, 10, 20, and 25 iterations shown in the figure consecutively. The resulting output is presented at the bottom of the corresponding mask to show the size reduction of the generated product in the output image.

5. Experimental Results

Experiments were conducted to evaluate the performance of the proposed VPP system. For these experiments, five sample images of an "Amazon Echo Dot" were chosen. $1,000$ augmented images of each product created from these five sample images were used to fine-tune the DreamBooth model using the text prompt "A photorealistic image of a sks Amazon Alexa device." The model was fine-tuned for $1,600$ steps, employing a learning rate of 5×10^{-6}, and a batch size of 1.

The fine-tuned model can inpaint products into the masked region. However, issues such as lack of product appearance, poor resolution, and disproportionate shape persist. The goal of the proposed Alignment Module is to automatically detect these issues. If identified, the problematic images are discarded, and a new image is generated from different random noise. Only if a generated image meets all the module's criteria it is presented to the user. Otherwise, a new image generation process is initiated. This loop continues for a maximum of 10 iterations.

5.1. Assessing Alignment Module

To assess the effectiveness of the Alignment Module, images were generated both with and without it. For each submodule, as well as for the overall Alignment Module, 200 images were generated: 100 with the filter activated and 100 without (referred to as the "Naive" case).

To prevent bias, all images were given random names and were consolidated into a single folder. These images were also independently evaluated by a human, whose scores served as the ground truth. This ground truth information was saved in a separate file for the final evaluation, which followed a blindfolded scoring method. All the experiments were also repeated for another product named "Lupure Vitamin C".

5.2. Evaluation Metrics

The evaluation and scoring method of each of the submodules of the Alignment module is described in the consecutive segments.

- **Content Score** For the image content score, images are categorized into two classes: 'success' if the product appears, and 'failure' otherwise. When the content module is utilized, the Failure Rate (FR), defined as the ratio of Failure to Success, is below 10% for both of the products.
- **Quality Score** For the quality score, images are rated on a scale from 0 to 10: 0 indicates the absence of a product, and 10 signifies a perfect-looking product. To evaluate in conjunction with the CLIP score, both the Mean Assigned Quality Score (MAQS) and Mean Quality Score (MQS) are calculated. MAQS represents the average score of images labeled between 0 and 10, while MQS is the output from the quality module, essentially reflecting cosine similarity.
- **Volume Score** For the volume module, images are also rated on a scale from 0 to 10: 0 for a highly unrealistic size, and 10 for a perfect size representation. When evaluating the volume module, the content module is not utilized. Since the size score necessitates the presence of a product, images without any product are excluded from this evaluation. To gauge performance, the Mean Assigned Size Score (MASS) is calculated in addition to the CLIP score.

5.2.1 Overall Results

The results of individual evaluations are presented in Table 1. It can be observed from this table that using any of the sub-modules consistently produced better outcomes compared to when no filtering was applied across various metrics. The results of the comprehensive evaluation, encompassing all sub-modules, can be found in Table 2.

Table 1. Individual evaluation of content, quality, and volume sub-modules within the overall Alignment Module. "Naive" represents the outputs without any filtering sub-modules. *Content* classifies the presence of the product in the generated images. *Quality* measures the proximity of the generated product to the sample product images used to fine-tune the diffusion model. Finally, *Volume* identifies the size category of the product.

		Naive	**Content**
Amazon Echo Dot	Success	72	94
	Failure	28	6
	FR	38.89%	6.38%
Lupure Vitamin C	Success	87	100
	Failure	13	0
	FR	14.94%	0.0%

	Naive	**Quality**
CLIP	32.49 ± 3.69	33.80 ± 2.69
MAQS	4.41 ± 3.23	6.41 ± 1.90
MQS	0.75 ± 0.14	0.83 ± 0.06
CLIP	24.61 ± 2.4	25.23 ± 2.66
MAQS	5.65 ± 2.85	6.47 ± 1.09
MQS	0.81 ± 0.13	0.86 ± 0.04

	Naive	**Volume**
CLIP	32.58 ± 3.70	33.42 ± 2.69
MASS	3.01 ± 2.68	4.81 ± 2.31
CLIP	24.22 ± 3.01	24.51 ± 2.89
MASS	5.64 ± 3.05	7.14 ± 1.53

Table 2. Comparison of the proposed method with and without using the Alignment Module in addition to the Paint-By-Example (PBE) [23] inpainting model. The "Naive" performance represents the generated output without applying the Alignment Module. The "Alignment" column represents the generated outputs where three cascaded filtering sub-modules are used, i.e., the Alignment Module.

	Amazon Echo Dot			Lupure Vitamin C		
	PBE	Naive	**Alignment**	PBE	Naive	**Alignment**
CLIP	31.44 ± 3.43	32.85 ± 3.19	33.85 ± 2.54	27.01 ± 2.10	24.71 ± 2.64	24.89 ± 2.90
MAQS	1.13 ± 1.30	4.65 ± 3.60	6.31 ± 2.39	1.75 ± 1.51	6.60 ± 3.01	7.81 ± 1.13
MASS	1.22 ± 1.60	3.05 ± 2.98	4.70 ± 2.81	2.43 ± 2.07	6.25 ± 3.08	7.30 ± 1.59
MQS	0.64 ± 0.08	0.75 ± 0.14	0.82 ± 0.05	0.67 ± 0.06	0.82 ± 0.12	0.86 ± 0.05
FR	78.57%	29.87%	0.00%	38.89%	17.64%	0.00%

(a) (b) (c) (d)

Figure 5. Inpainted product image of Paint-by-Example (PBE). PBE generates high-quality images which explains the higher CLIP score in the case of Lupure Vitamin C. However, the inpainted product does not look similar to the desired product at all resulting in very poor mean assigned quality and size scores. Output images for Amazon Echo Dot is shown in (a) and (b), and for Lupure Vitamin C is shown in (c) and (d).

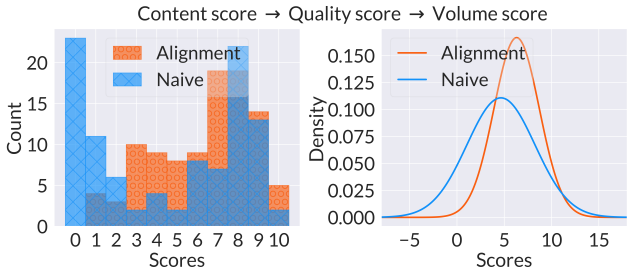


Figure 6. Empirical performance of Alignment Module for Amazon Echo Dot. Noticeably, no output is generated without any product when the Alignment Module is employed. Moreover, the mean quality score has increased from 4.65 to 6.31.

5.3. Comparison with Paint-By-Example

The proposed method is compared with the Paint-By-Example (PBE) [23] inpainting model and Table 2 shows the performance comparison of the proposed method along with PBE. PBE can generate very high-quality images, however, the inpainted product in the generated image does not look alike the desired product at all as shown in Figure 5 resulting in very poor MAQS and MASS. Whereas the inpainted product of our proposed method resembles much of the original product shown in Figure Figure 7.

5.4. Frequency Distribution

The frequency distribution and density function of the assigned quality scores in the case of "Naive" and "Alignment" for Amazon Echo Dot is presented in Figure 6. The density mean has shifted from 4.65 to 6.31 when Alignment Module is adopted indicating the effectiveness of the proposed module.

6. Path to Production

6.1. Product API

The location identifier, fine-tuned model, and Alignment Module are combined to develop an easy-to-use VPP Streamlit web app [1]. This app is hosted on Amazon Sagemaker using an "ml.p3.2xlarge" instance, which is a single V100 GPU with 16GB of GPU memory. The demo app's interface is illustrated in Figure 8. In the top-left 'Image' section, users can either upload their own background image or choose from a selection of sample background images to generate an inpainted product image.

The web app provides extensive flexibility for tuning the parameters of the Alignment Module so that users can comprehend the effects of these parameters. In the 'seed' text box, a value can be input to control the system output. The segmentation threshold for CLIPSeg defaults to 0.7, but users can refine this value using a slider. Within the 'Mask Params' section, the number of dilation and erosion iterations can be set and visualized in real-time.

The filter, represented by the Alignment Module, can be toggled on or off. The 'Max Attempt' slider determines the number of regeneration attempts if the model doesn't produce a satisfactory output. However, if a seed value is specified, the model will generate the output only once, regardless of the set value. Lastly, in the 'Filter Params' section, users can fine-tune the threshold values for each sub-module of the Alignment Module, specifically for content, quality, and volume.

The "show stats" button beneath the input image displays the mask alongside details of the model outputs. These details include the seed value, placement, generated and modified captions, and the content, quality, and volume/size scores. By visualizing the mask and its area, users can apply erosion or dilation to adjust the product's size. The default threshold values for content, quality, and volume are 0.7, 0.7, and 0.34, respectively. While these values can be adjusted slightly higher, it's recommended to also set the 'Max Attempt' to 10 in such cases. A higher threshold means that the generated output is more likely to fail the criteria set by the Alignment Module.

[1] STREAMLIT: `https://streamlit.io/`

6.2. Future Considerations for Product Scalability

Fine-tuning stable diffusion using DreamBooth can take up to 30 minutes, depending on dataset size, image resolution, and extent of training. When considering a customer with hundreds or thousands of products, this process could take days to complete model training across different products.

Our pipeline is deployed on Amazon SageMaker, a managed service that supports the automatic scaling of deployed endpoints. This service can dynamically accommodate large computational needs by provisioning additional instances as required. As such, fine-tuning 100 SD models for 100 different products would still only take about 30 minutes if 100 instances were utilized in parallel.

The fine-tuned models are stored in an Amazon S3 (Simple Storage Service) bucket, with each model being 2.2 GB in size. Consequently, 100 fine-tuned models would occupy approximately 220 GB of storage space. A pertinent question arises: Can we strike a space-time trade-off by training a single model with a unique identifier for each product?

If this is feasible, the space requirement would be reduced to a consistent 2.2 GB. However, that one model would need more extensive training - specifically training steps would increase by a factor of 100 for 100 products, thereby lengthening the computation time. This approach remains untested and warrants future exploration [10].

7. Conclusion

In this paper, we present a novel, fully automated, end-to-end pipeline for Virtual Product Placement. The proposed method automatically determines a suitable location for product placement into a background image, performs product inpainting, and finally evaluates image quality to ensure only high-quality images are presented for the downstream task.

Using two different example products, experiments were conducted to evaluate the effectiveness of the proposed pipeline, the performance of the individual sub-modules, and the overarching Alignment Module. Notably, when upon employing the Alignment Module, the Failure Ratio (FR) plummeted down to 0.0% for both investigated products. Additionally, images produced with the Alignment Module achieved superior CLIP, quality, and size scores.

Qualitatively, the produced images present a clean and natural semantic inpainting of the product within the background image. The accompanying web application facilitates pipeline deployment by enabling image generation through a user-friendly interface with extensive image fine-tuning capabilities. The high-quality integration of products into images underscores the potential of the proposed VPP in the realms of digital marketing and advertising.

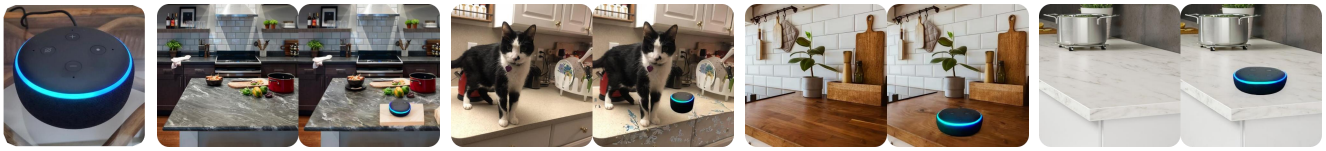


Figure 7. Qualitative results of the proposed VPP system. Experiments are performed using two different products, Amazon Echo Dot as shown on top, and Lupure Vitamin C as shown on bottom. The original training images are shown on the left, and then the pairs of background and inpainted output images are presented side by side.

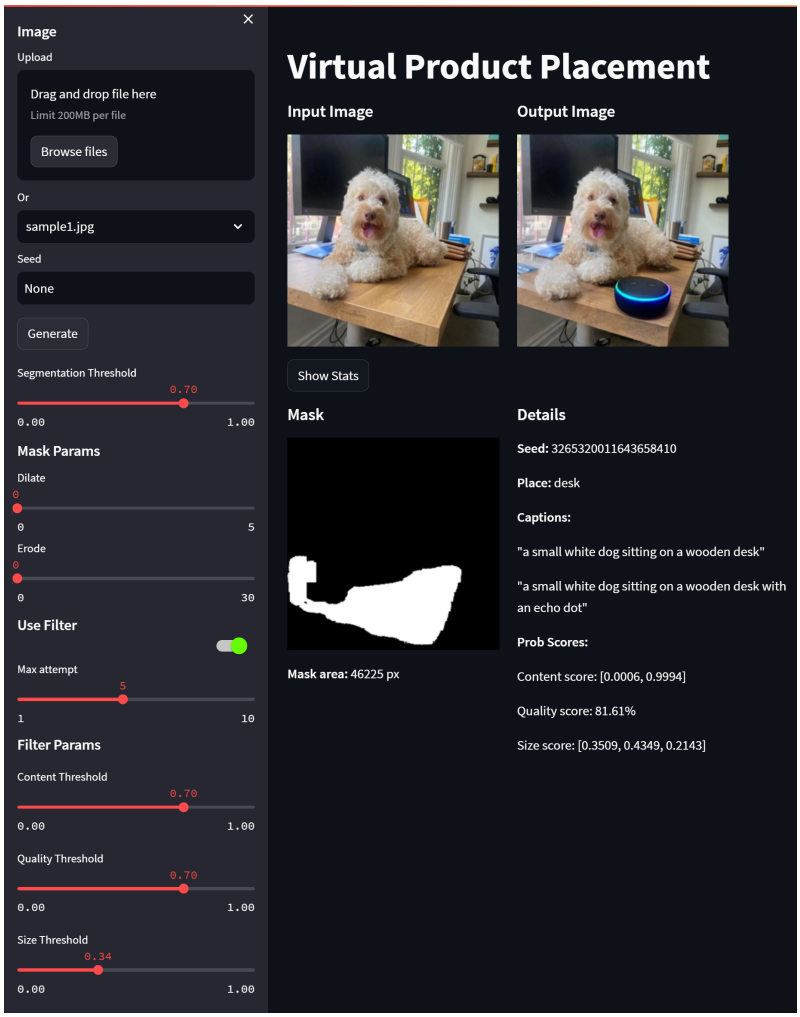


Figure 8. The interface of the VPP web app demo was built using Streamlit hosted in Amazon SageMaker. The uploaded background image is shown under the title "Input Image" and the inpainting image with an Amazon Echo Dot is shown under the title "Output Image". Moreover, the generated mask produced by the location identifier and the other intermediate details of the proposed VPP system is also presented in the interface.

References

[1] 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. 1, 2

[2] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021. 2

[3] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2

[4] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018. 2

[5] Zachary Glass. The effectiveness of product placement in video games. *Journal of Interactive Advertising*, 8(1):23–32, 2007. 1

[6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[8] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 1, 2

[9] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 3

[10] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 7

[11] 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. 1, 3

[12] Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicmix: Semantic mixing with diffusion models. *arXiv preprint arXiv:2210.16056*, 2022. 2

[13] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 4

[14] Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. A frustratingly simple approach for end-to-end image captioning. *arXiv preprint arXiv:2201.12723*, 2022. 4

[15] John McDonnell and Judy Drennan. Virtual product placement as a new approach to measure effectiveness of placements. *Journal of Promotion Management*, 16(1-2):25–38, 2010. 1

[16] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4

[18] 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. 1

[19] 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. 2, 3

[20] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 3

[21] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[22] Ye Wang and Huan Chen. The influence of dialogic engagement and prominence on visual product placement in virtual reality videos. *Journal of Business Research*, 100:493–502, 2019. 1

[23] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 3, 6