Concept-Edge Fusion: Background Generation for Product Presentation Based on Text-to-Image Model

Pengfei Deng[1,2] Tianjiao Zhang[1] Weize Quan[1,2*] Hanyu Wang[3] Qinglin Lu[4]
Zhifeng Li[4] Dong-Ming Yan[1,2]
1. MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.
qweizework@gmail.com (*Corresponding author)
2. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China.
3. Department of Computer Science, University of Maryland, College Park, MD 20742, USA.
4. Tencent, Shenzhen 518057, China.

Abstract

This work presents a Text-to-Image model-based product image background generation method, Concept-Edge Fusion, which generates a high-quality background for a product freely through text description while maintaining the details of the product and without deforming the edges. Existing methods for generating product backgrounds often lead to semantic misunderstanding and edge expansion, which greatly undermines the quality of the generated image. Semantic misunderstanding represents the apparent difference between the main subject of the generated image and that of the reference product image when the model generates a new image based on the text description. Edge expansion refers to the apparent changes in the contour/shape of the input product. To solve these problems, we introduce Concept-Inject and Edge-Control modules to help the Text-to-Image model better generate the background guided by the text description. The concept-inject module prevents the model from semantic misunderstanding about the given product, and the edge-control module ensures that the product edges are not expanded when completing the product background. Extensive experiments demonstrate that our method can better perform background generation for products without changing the semantics, shape, and details of the original products. We also construct a dataset to evaluate the text-guided product background generation.

Keywords: Text-to-image model, Background generation, Subject concept injection, Edge control, Product

1. Introduction

With the advance of diffusion models [31, 16], image generation has achieved rapid development in recent years, especially in the field of text-to-image generation[28, 25, 27, 30, 2]. Taking DALL·E 3 [1] and Midjourney [2] as examples, they have demonstrated remarkable capabilities in generating photorealistic images that closely align with textual inputs. To explore the commercial application of text-to-image generation, many methods [8, 4, 1, 21, 22] have made excellent attempts based on text-to-image models. They aim to convert text descriptions into controllable, high-quality, realistic images to meet various commercial application needs. These methods can rapidly generate a wide range of image types, *e.g.*, product posters, advertising designs, and film special effects, and thus enhance production efficiency.

Figure 1. Illustration of semantic misunderstanding and edge expansion. Prompt: "A jar On the calm lake, the mountain and mist is in the distance, with the mountain reflecting on the lake, professional photography, and commercial photography."

In this paper, we focus on keeping the subject of the product while generating the background based on the text, ensuring that the product and background blend in as naturally as possible. By providing only the reference image of the product and its mask, we can generate the desired background based on the text. This potential capacity holds the promise of facilitating text-to-image applications in e-commerce contexts, thereby permitting the personalization

[1] https://openai.com/dall-e-3
[2] https://www.midjourney.com/home

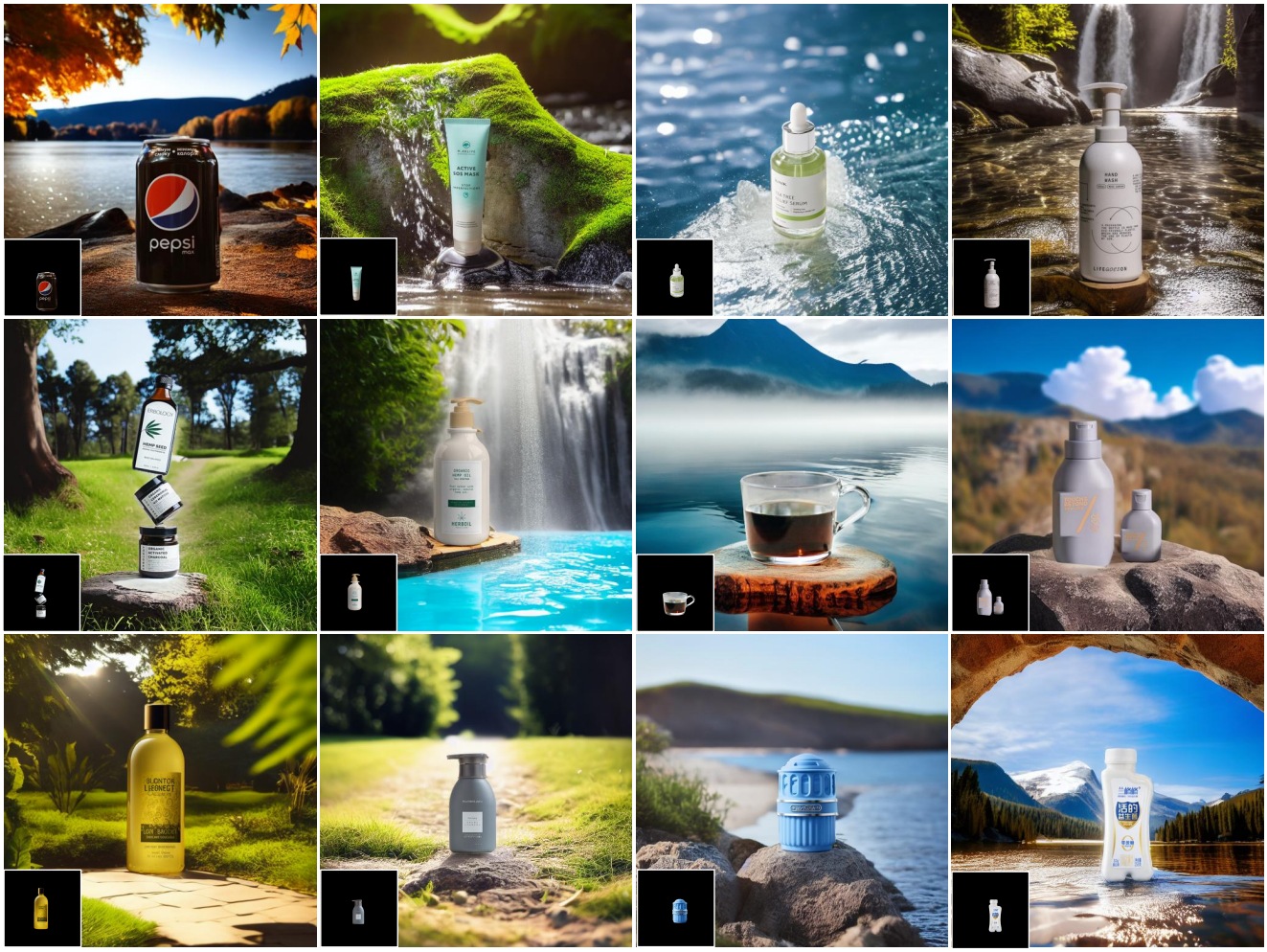

Figure 2. Some results of our proposed Concept-Edge Fusion. The backgrounds of all products are generated from text descriptions, and there are no errors in product recognition and edges.

of product backdrops. Consequently, this would foster the pragmatic incorporation of text-to-image paradigms in real-world environments.

Despite the urgent need, previous researchers have not explored this topic well. To implement the function of changing the background for specific products, most researchers used customized image generation methods [29, 8, 12, 39, 3, 44] or image composition techniques [47, 32]. Although these methods often maintain consistent primary features across diverse backgrounds for the product, they all change the product's detailed information, such as text, logo, texture, etc. In contrast, the methods based on the Stable Diffusion [28] inpainting model and the ControlNet [48] inpainting model can achieve fewer changes in the details within the product area. However, they often encounter problems such as semantic misunderstanding and edge expansion, as shown in Fig. 1, which affects the harmonization of the subject and the background and the beauty of the generated images.

To address the above issues, we propose a novel method called "Concept-Edge Fusion", which consists of two main modules, *i.e.*, the Concept-Inject module and the Edge-Control module. The Concept-Inject module can inject the semantic concept of the product into the base model, allowing the base model to know that the product described in the text is consistent with the given reference product image. The Edge-Control module can not only inject the subject feature information but also provide fine-grained edge control to ensure that the edges of the product do not expand. With these two carefully designed components, our method achieves superior generation performance as shown in Fig. 2.

In addition, there is no public standard dataset for evaluating text-guided product background generation. Therefore, we propose a curated evaluation dataset (*i.e.*, ProductBench) that captures a wide variety of product and textual-

description backgrounds. Each sample consists of (i) original product image, (ii) product mask, and (iii) background textual description. By generating a batch of product background images on the test set, we can more accurately assess the model's ability to perform background generation for a wide range of products.

2. Related Work

2.1. Text-guided Image Inpainting

Some methods were attempting to generate content within the mask area using text guidance [26]. DiffEdit [11] utilized adaptive methods to obtain the mask of the editing area and add noise to the input image using DDIM. During the denoising process, the content within the mask area is generated using text descriptions. Repaint [23] used a pre-trained DDPM [16] model to achieve inpainting. In each step, it samples the known region from the input and the inpainted part from the DDPM output. Blended Diffusion [1] conducted multi-step blending in the masked region to generate more harmonized outputs. Paint-by-example [43] applied mask shape-augmentation and reference-augmentation to image inpainting. AnyDoor [8] achieved subject generation at any position with image condition and ControlNet based on high-frequency information. Smart-Brush [41] generated objects by incorporating both text and shape guidance with precision control. [35] introduced an edge predictor to guide the diffusion model, ensuring that the edges of the synthesized image align with an input sketch. [34] proposed an edge-preserving diffusion process to enhance the structural details. Inpainting-Anything [46] involved SAM [18] and SD [28] to replace any object in the source image with the text-described target. However, these existing methods often generate an unharmonious background for a given product and SD-based methods struggle to preserve the original input product.

2.2. Customized Image Generation

Customized (subject-driven) generation uses several reference images to learn the given subject. Some works [7, 29, 24, 36, 13] learned the given reference images by learning new vocabulary. In addition, some methods [12, 39, 17, 44, 8, 43] used image encoders to directly inject the concepts of reference images into the UNet model, eliminating the need to learn new vocabulary for new reference images. Other methods [4, 21, 22, 40, 15] achieved multi-subject generation by editing the cross-attention map of the image. Although these methods have achieved decent personalized editing effects to some extent, they all come with certain limitations. For instance, learning a new concept requires a considerable amount of time to fine-tune the model. In addition, although the image encoder-based approach can directly introduce concepts into the model, it cannot guarantee that the details of the subject are completely preserved.

2.3. Image Harmonization

A classical image composition pipeline is cutting the foreground object and pasting it on the given background. Image harmonization [33, 6, 42, 5, 10, 9, 14] could further adjust the pasted region for more reasonable lighting and color. However, these methods only explore the low-level changes, such as editing the structure, view, and pose of the foreground objects, or generating the shadows and reflections. Recently, ObjectStitch [32] and ControlCom [47] utilized diffusion models to achieve object composition. Although they can blend the subjects into the given background, the subjects are changed and the subjects' details are not well-maintained. The color, texture, and shape of the object may undergo certain changes, making it unsuitable for background generation of the product.

3. Method

Our proposed Concept-Edge Fusion pipeline is demonstrated in Fig. 3. Given a reference product, a product mask, and a background description, the goal is to place this product in any background based on the textual description and present a natural blend between the product and the background. Our framework mainly contains three modules: the Concept-Inject module, the Edge-Control module, and the pre-trained SD [28] model. To effectively modify the background of the product, we use product data for training the Concept-Inject module and Edge-Control module, and these two modules are trained independently. The loss function remains consistent with that of training the diffusion model.

3.1. Concept-Inject Module

We propose Concept-Inject module to enhance the base model's perception for the provided product subject in the reference image. This module allows noisy image features in the UNet to interact with the features of the reference image using newly designed cross-attention layers.

Specifically, we use a pre-trained CLIP image encoder to extract the features of the reference image, resulting in a sequence of image embeddings. This sequence undergoes processing via the object mask, ensuring that it is localized within the area of the object, which facilitates more accurate removal of the reference image's background and improves the integration of the reference object with the background described in the text.

Then, to enable a more effective adaptation to the inherent characteristics of the reference image, a Mapper network is employed to reduce the length of the image embedding sequence to N (we set $N = 16$ in this study). Consequently, the shape of the final reference embedding se-

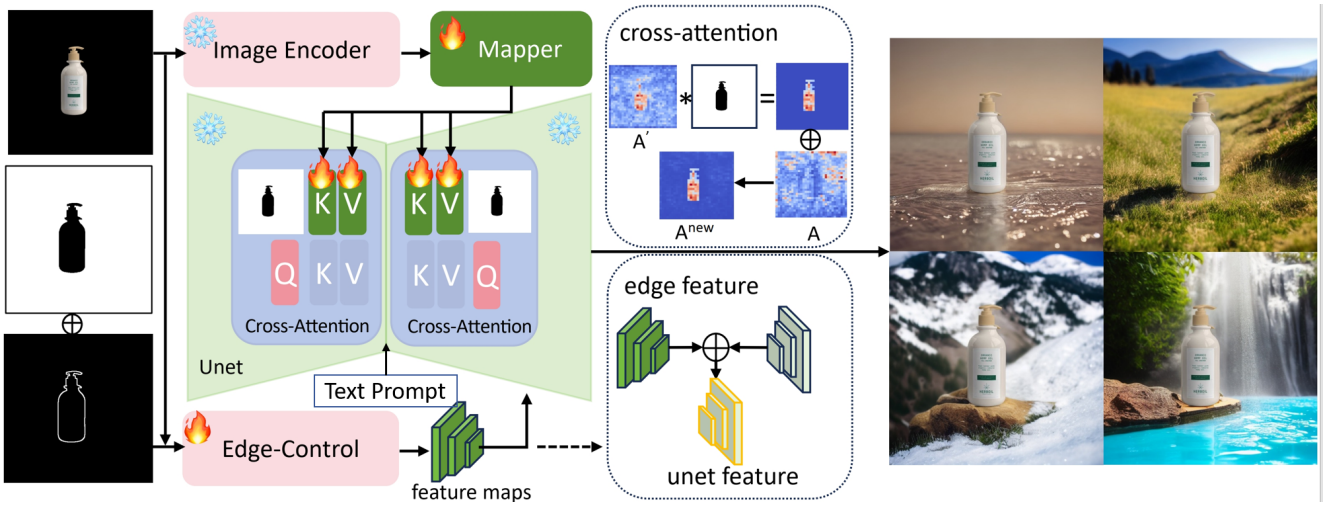

Figure 3. Overall pipeline of Concept-Edge Fusion, which is designed to generate product backgrounds through text descriptions. The pipeline mainly consists of three components: Concept-Inject, Edge-Control, and SD model. Concept-Inject module is composed of an image encoder of CLIP, Mapper, and cross-attention layer of UNet with K and V linear layers that are updated during training. The Mapper is composed of MLP and responsible for mapping the image embedding to a sequence embedding suitable for UNet. The cross-attention calculated from the image utilizes the mask to extract attention within the region, reducing the influence of the reference image on background generation and enhancing the concept preservation as well. Edge-Control is a ControlNet-like module, with inputs being the masked image, mask, and canny edge, providing better edge control. **Flames** and **snowflakes** represent learnable and frozen parameters, respectively.

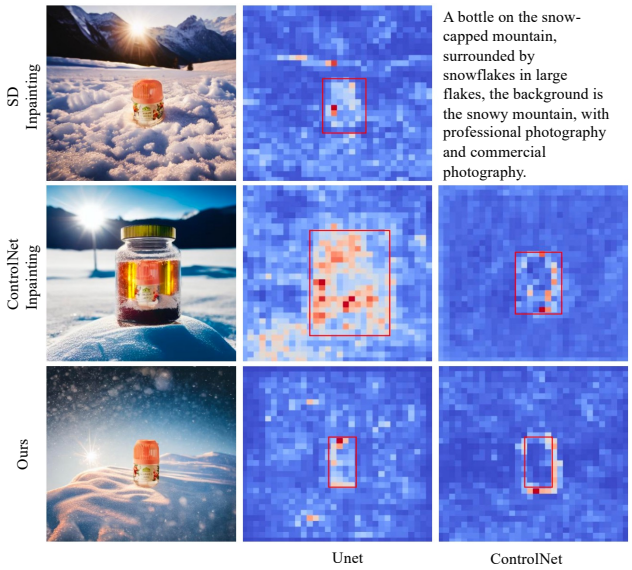

Figure 4. Attention map of the cross-attention layer. The resolution of the attention map is 32×32.

quence is (B, N, D), where B represents the batch size, N is the length of the embedding sequence, and D represents the dimension of the embedding.

Subsequently, through a series of added cross-attention layers, the reference embedding sequence is injected into the original cross-attention layers of the UNet model via a process of weighted summation. Mathemetically, given the original cross-attention $\mathbf{A}$, the calculation method for the added cross-attention $\mathbf{A}^{'}$ of injected embedding sequence is as follows:

$$\mathbf{A}^{'} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K_c}^{\top}}{\sqrt{d}}\right)\mathbf{V_c}, \tag{1}$$

where $\mathbf{Q} = \mathbf{X}\mathbf{W_q}$, $\mathbf{K_c} = \mathbf{T}\mathbf{W_k}$, $\mathbf{V_c} = \mathbf{T}\mathbf{W_v}$, $\mathbf{W_k}$ and $\mathbf{W_v}$ are copied from the original cross-attention layer and updated during training to adapt to the reference image features, $\mathbf{W_q}$ is the original Q linear layer in the cross-modal attention layer of UNet, $\mathbf{T}$ is the feature embeddings obtained from the reference product image through Mapper, and $\mathbf{X}$ is the feature embeddings of the noisy image processed by the UNet network. In the training stage, the original UNet model is frozen. The calculation mechanism for the new cross-attention becomes:

$$\mathbf{A}^{new} = \mathbf{A} + \lambda(\mathbf{A}^{'} * mask), \tag{2}$$

where λ is a control coefficient and $mask$ is object mask, new cross-attention $\mathbf{A}^{new}$ is the cross-attention of the UNet in our pipeline.

As shown in Fig. 4, we visualize the cross-attention maps of the ControlNet inpainting and SD inpainting. From the figure, we can see that both methods have shortcomings in subject recognition and edge control. The concept recognition based on ControlNet inpainting has more serious problems. The reason is that during the denoising process, Con-

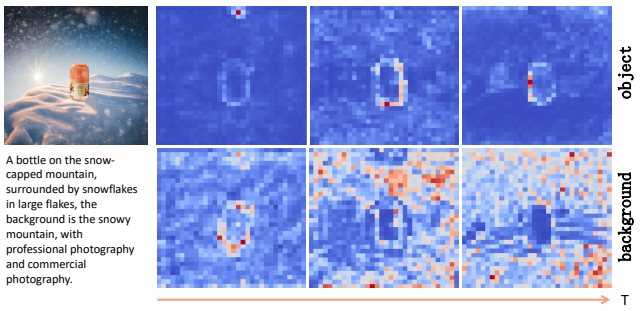

Figure 5. Cross-attention map of Edge-Control module. The prompt for generating the image is: " A jar placed in an outdoor snowy scene." The attention map contains the object token " jar " and the background token " outdoor ". As the iterations of denoising, the Edge-Control module gradually distinguishes between the background and the object.

trolNet only adds control condition information to the decoder of UNet, but the encoder of UNet mainly generates content based on text. If the encoder generates content with apparent recognition errors for the subject, it will cause the failure of the control conditions of the ControlNet part. Due to the lack of an efficient subject injection mechanism and edge condition control, SD inpainting also exhibits the input product's change.

3.2. Edge-Control Module

In the process of generating the background of a product, it is imperative to ensure that the edges of the product are not distorted. In this regard, we propose the Edge-Control module, a ControlNet-style module, which serves as an effective means of achieving the desired contour control. To further enhance both the object recognition capabilities and edge control, we incorporate three distinct conditional information: the masked product image, the product mask, and the canny of the product mask. By jointly utilizing these three elements, we aim to optimize the overall performance of the system in terms of accurately identifying objects and maintaining precise control over their shapes.

Technically, they are concatenated together in the channel dimension. The masked object image is used to enhance the injection of object features, the mask provides shape perception, and the canny is used to further enhance the control of edges. By incorporating edge and conceptual features at each resolution level within every block of the Edge-Control module, the perception of conceptual and edge information in the UNet encoder is effectively enhanced. A gradual improvement in object-background fusion is achieved through iterative refinement across time steps. Fig. 5 illustrates the changes of the attention map in the edge-control module with denoising time steps. It can be seen that the distinction between the object and the background becomes more and more obvious.

3.3. Training Strategies

Product Image. Inspired by the previous work [38, 41], shape guidance can achieve precise inpainting control. To achieve better control over the shape of the product, we used the product data when training the two modules. All images are high-quality product photography data downloaded from Pinterest [3], in total 100,000 images. The images are automatically annotated with captions using LLaVA-v1.5-13B [20], and the corresponding product areas are cropped out for the mask and mask canny edges.

Random Mask. To enhance the model's perception ability in foreground and background recognition, we have adopted some special strategies during the training process of the Edge-Control module. Specifically, we flip the mask with a 25% probability, which can increase the model's adaptability to different scenarios and improve its accuracy in processing complex images. At the same time, we also leave the text blank with a 50% probability. This operation aims to reduce the interference of text description on the model's judgment, thereby improving the model's ability in conditional image control. By combining these two strategies, the Edge-Control module can better distinguish foreground and background elements in practical applications. Moreover, the quality and effect of image processing can be remarkably improved.

4. Experiments

4.1. Implementation Details

Network Settings. We choose Stable Diffusion V1.5 [28] as the base model. The Mapper is a single MLP with LayerNorm. The architecture of the Edge-Control module is similar to ControlNet with 9-channel input. During the training process, the Edge-Control module and Concept-Inject module are trained separately, allowing for optimization tailored to their respective characteristics, We did not use data to train the base model. In terms of input images, we use a uniform resolution of 512×512 to ensure consistency in the training data. For images with an aspect ratio not equal to 1, we adjust the image and mask to square images by filling it with pure white to meet the model's input requirements. When processing the masked image tensor, we first normalize the original image and then use the mask to obtain the masked image tensor. We choose the AdamW optimizer with an initial learning rate of 1e-5 and warmup 500 steps, during the inference stage λ is set to 0.5.

Benchmarks. Since there is no available benchmark for evaluating product background generation, we propose a new benchmark, which includes 37 different products. These 37 products have not appeared in the training dataset

[3] https://www.pinterest.com

Figure 6. Comparisons with state-of-the-art methods. The first column is prompt, and the remaining columns are the generated images by different methods.

and each has 30 different backgrounds. Therefore, we collect a total of 1,110 images for comparison purposes.

Evaluation Metrics. We employ the CLIP Score and PickScore [19] to assess the consistency between text and images automatically. We use PSNR and SSIM to measure the preservation of input products. To evaluate the aesthetic appeal of images, we utilize the SAAN [45] model to calculate aesthetic scores. Moreover, we have organized user studies involving a group of 15 annotators to assess the generated results in terms of concept, edge, and aesthetics.

4.2. Comparisons with State-of-the-art Methods

Qualitative comparisons. In Fig. 6, we have demonstrated a comparison of our method with open-source SOTA methods, *i.e.*, Blended diffusion [1], Magic [37], ControlNet Inpainting and Stable Diffusion Inpainting. We maintain the same input subject and prompt for all methods. The background generated by Magic is not harmonious with the

Figure 7. Multi-object background generation. Our method can correctly identify both objects.

Table 1. Quantitative analysis. We use CLIP, SAAN [45] and Pickscore [19] to quantitatively evaluate the semantic alignment and visual quality. We utilize PSNR and SSIM to assess product preservation in unmasked regions. Higher is better.

Method	CLIP	Aesth.	Pickscore	PSNR	SSIM
ControlNet Inpainting	35.25	4.63	19.19	21.68	0.7159
SD Inpainting	35.29	4.69	19.32	22.55	0.7069
Blended diffusion	28.16	4.72	18.70	24.00	0.7980
Magic	27.09	4.61	17.86	22.64	0.7687
Our	**35.33**	**4.78**	**19.42**	**48.28**	**0.9734**

input product. The results of Blended diffusion, ControlNet Inpainting, and SD Inpainting often produce subject concept errors and shape change. In contrast, our method can correctly recognize the subject and effectively control edge information. Among all methods, our method better preserves the original details, *e.g.*, the words and logos. This superior performance is attributed to our proposed concept-inject and edge-control modules.

In addition, we demonstrate the comparison of generated images for multiple subjects, and the corresponding results are shown in Fig. 7. Based on the experimental results, the background generated by Blended diffusion and Magic are not very consistent with the text description about expired background. ControlNet Inpainting and SD Inpainting both exhibit a certain degree of cognitive bias towards multiple objects; ControlNet Inpainting has a greater bias. In some cases, ControlNet Inpainting may perceive the lamp as a bottle, and SD Inpainting may recognize the bottle as a different type of bottle. Our method can exactly recognize each subject and generate reasonable backgrounds.

Quantitative comparisons. For generated images of various objects with different backgrounds, we employ the PSNR, SSIM, CLIP Score, PickScore, and Aesthetic Score as metrics to evaluate the performance of multiple methods in terms of product preservation, conceptual understanding, and the aesthetic quality of object-background fusion. The corresponding results are reported in Table 1. Among all methods, Magic has the weakest perception of the object concept and aesthetics, while our method achieves the best cognition and aesthetic scores. This means that the generated background by our method has better consistency with the text description and higher visual quality. In terms of PSNR and SSIM, our method achieves the best performance, which illustrates the superior preservation ability of input products. These results are also consistent with the qualitative comparisons as shown in Fig. 6 and 7.

User study. We conduct a user study to further compare the results of ControlNet Inpainting, SD Inpainting, Blended

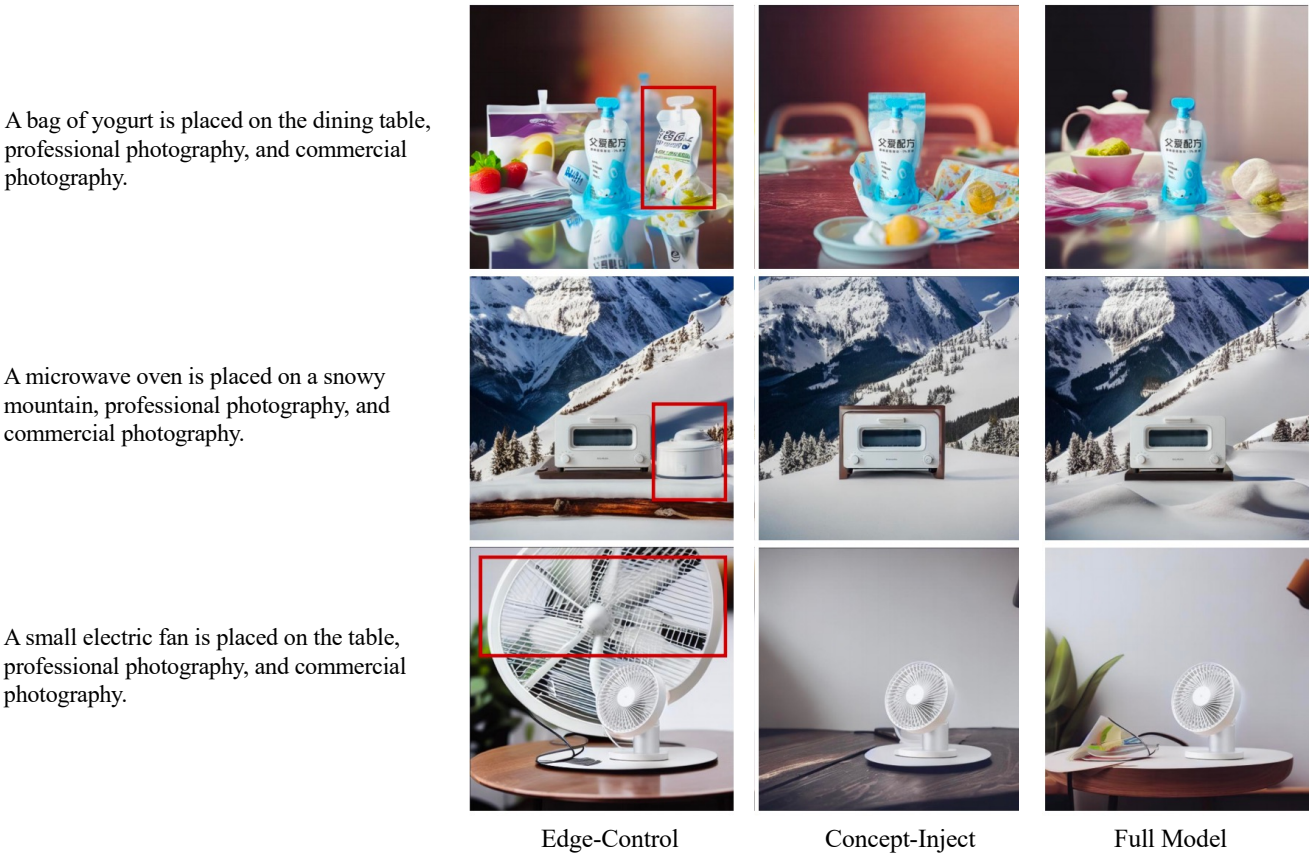

Figure 8. Qualitative ablation studies on Concept-Inject and Edge-Control modules of our method. Using only Edge-Control may result in object recognition errors without exact concept guidance, while using only the Concept-Inject module cannot precisely control the object's edge. Combining both modules enables accurate object recognition and precise edge control.

Table 2. User study. Evaluation metrics include "Recognition", "Boundary", and "Aesthetics". The scoring range for each metric is from 1 (worst) to 4 (best). Our method achieved the highest human evaluation scores compared to other advanced methods.

Method	Recog.	Bound.	Aesth.
ControlNet Inpainting	1.94	2.06	1.82
SD Inpainting	2.58	2.78	2.71
Blended diffusion	1.51	2.09	1.37
Magic	1.35	2.41	1.39
Our	**3.25**	**3.59**	**3.36**

diffusion, Magic, and our method. Specifically, we ask 10 participants to judge 400 sets of image comparisons (randomly selected from all test results). These participants possess fundamental image processing skills. For each set, a corresponding product object image is provided as a reference. We conduct a detailed evaluation from three perspectives: "Recognition" (object recognition ability), "Boundary" (whether the boundary expands), and "Aesthetics" (aesthetic blending of the object and background). We prepare detailed regulations and templates to rate the images with scores of 1 to 4 for the three perspectives. The results of our user study are reported in Table 2. We can see that Blended diffusion and Magic have the poorest concept recognition and edge control, leading to a decline in the overall aesthetics of the generated images. SD Inpainting has relatively better concept recognition and edge control. Among all methods, our method has the best concept recognition and fusion of the object and background, resulting in the highest aesthetic score accordingly.

Table 3. Quantitative ablation study. We quantitatively evaluate the impact of only using a single module on the model's conceptual awareness and edge control capabilities. Higher is better.

Method	CLIP	Aesth.
Edge-Control	30.53	4.72
Concept-Inject	30.56	4.74
Our	**35.33**	**4.78**

4.3. Ablation Studies

In this part, we carry out extensive ablation studies to verify the effectiveness of our designs. To evaluate the ra-

w/o mask mask w/o mask mask

Figure 9. Qualitative ablation studies on the masking mechanism in Concept-Inject module. "w/o mask" will interfere with background generation.

Figure 10. Results of our proposed Concept-Edge Fusion. The same product can naturally blend in different backgrounds.

tionality of the model design, we separately utilize each module to generate images as shown in Fig. 8. It becomes evident that relying on a single module results in insufficient control capabilities, leading to conceptual cognition errors and edge expansion phenomena. By contrast, employing two modules simultaneously allows for a collaborative enhancement of concept perception and edge control. The Concept-Inject module can enable the model to correctly perceive the correspondence between the given object and the object described in the text when generating the background, thus preventing the generation of cluttered objects introduced by the text in other areas of the background. The Edge-Control module, on the other hand, can help with better edge control, thereby achieving more accurate object recognition. When relying solely on text prompt and Edge-Control module for attention computation can lead to in-

Figure 11. Illustration of out-of-domain generalization, where all products are unseen during the training.

accuracies, potentially causing semantic misunderstanding (e.g., generating other related objects). Numerical results are listed in Table 3, which are aligned with our visual analysis. Using the Concept-Inject and Edge-Control modules separately will weaken the model's capabilities of conceptual recognition and edge control.

In Fig. 9, we explore whether to apply a masking mechanism for the attention calculation in the Concept-Inject module (Eqn. (2)). If the mask is not used to extract the object's attention in the Concept-Inject module, it will lead to the interference of background information with masked images in the attention. Because CLIP encodes background information along with the image when converting it to an embedding, the token's output embedding contains not only object information but also background information, which in turn will cause the generated image background to be disturbed (the "w/o mask" of Fig. 9). Using the mask in the Concept-Inject module, the generated background is clear and natural, as shown in the "mask" of Fig. 9.

4.4. More Evaluation

To further demonstrate that our approach can not only position objects in a wide range of backgrounds but also maintain the ability to recognize objects and precisely control edges, we randomly selected an object from the test set and generated images in different scenes based on the different background descriptions. As can be seen in Fig. 10, the object can naturally blend with various types of backgrounds, where the edges of the object are precisely controlled and the generated background naturally blends with the light and shadow on the object's surface. In Fig. 11, we illustrate the background generation for various products that are not included in the training data. The generated images are visually pleasing. This implies that our method has the capacity of out-of-domain generalization.

5. Conclusion

In this work, we present Concept-Edge Fusion, implementing automatic product background generation based on diffusion models. The core contribution of our method is proposing the Concept-Inject and Edge-Control modules to accomplish the concept cognitive of given subjects and provide exact edge control, enabling high-quality text-guided background generation. This allows different products to naturally blend with various backgrounds, making it possible to quickly change the background of objects through text descriptions. We also collected a benchmark used to evaluate the performance of product background generation. Extensive experiments demonstrate the superiority of our method. In the future, we would like to enhance the lighting and shadows of product background generation.

Acknowledgement

This work was partially supported by the National Natural Science Foundation of China (62102418 and 62172415), the Beijing Science and Technology Plan Project (Z231100005923033), and the Excellent Youth Program of State Key Laboratory of Multimodal Artificial Intelligence Systems.

References

[1] O. Avrahami, O. Fried, and D. Lischinski. Blended latent diffusion. *ACM Transactions on Graphics*, 42(4):1–11, 2023. 1, 3, 6

[2] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 1

[3] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2

[4] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*, 42(4):1–10, 2023. 1, 3

[5] B.-C. Chen and A. Kae. Toward realistic image compositing with adversarial learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8407–8416, 2019. 3

[6] H. Chen, Z. Gu, Y. Li, J. Lan, C. Meng, W. Wang, and H. Li. Hierarchical dynamic image harmonization. In *ACM International Conference on Multimedia*, pages 1422–1430, 2023. 3

[7] H. Chen, Y. Zhang, S. Wu, X. Wang, X. Duan, Y. Zhou, and W. Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In *International Conference on Learning Representations*, 2023. 3

[8] X. Chen, L. Huang, Y. Liu, Y. Shen, D. Zhao, and H. Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023. 1, 2, 3

[9] W. Cong, X. Tao, L. Niu, J. Liang, X. Gao, Q. Sun, and L. Zhang. High-resolution image harmonization via collaborative dual transformations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18449–18458, 2022. 3

[10] W. Cong, J. Zhang, L. Niu, L. Liu, Z. Ling, W. Li, and L. Zhang. Dovenet: Deep image harmonization via domain verification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8391–8400, 2020. 3

[11] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations*, 2023. 3

[12] A. Elarabawy, H. Kamath, and S. Denton. Direct inversion: Optimization-free text-driven real image editing with diffusion models. *arXiv preprint arXiv:2211.07825*, 2022. 2, 3

[13] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023. 3

[14] Z. Guo, H. Zheng, Y. Jiang, Z. Gu, and B. Zheng. Intrinsic image harmonization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16362–16371, 2021. 3

[15] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023. 3

[16] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020. 1, 3

[17] X. Jia, Y. Zhao, K. C. Chan, Y. Li, H. Zhang, B. Gong, T. Hou, H. Wang, and Y.-C. Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 3

[18] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, pages 3992–4003, 2023. 3

[19] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. 2023. 6, 7

[20] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *Neural Information Processing Systems*, 2023. 5

[21] Z. Liu, R. Feng, K. Zhu, Y. Zhang, K. Zheng, Y. Liu, D. Zhao, J. Zhou, and Y. Cao. Cones: Concept neurons in diffusion models for customized generation. In *International Conference on Machine Learning*, volume 202, pages 21548–21566, 2023. 1, 3

[22] Z. Liu, Y. Zhang, Y. Shen, K. Zheng, K. Zhu, R. Feng, Y. Liu, D. Zhao, J. Zhou, and Y. Cao. Cones 2: Customizable image synthesis with multiple subjects. *arXiv preprint arXiv:2305.19327*, 2023. 1, 3

[23] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11451–11461, 2022. 3

[24] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 3

[25] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, volume 162, pages 16784–16804, 2022. 1

[26] W. Quan, J. Chen, Y. Liu, D.-M. Yan, and P. Wonka. Deep learning-based image and video inpainting: A survey. 132(7):2367–2400, 2024. 3

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

[28] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10674–10685, 2022. 1, 2, 3, 5

[29] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 3

[30] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, S. K. S. Ghasemipour, R. G. Lopes, B. K. Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Neural Information Processing Systems*, volume 35, pages 36479–36494, 2022. 1

[31] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, volume 37, pages 2256–2265, 2015. 1

[32] Y. Song, Z. Zhang, Z. Lin, S. Cohen, B. Price, J. Zhang, S. Y. Kim, and D. Aliaga. Objectstitch: Object compositing with diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18310–18319, 2023. 2, 3

[33] K. Sunkavalli, M. K. Johnson, W. Matusik, and H. Pfister. Multi-scale image harmonization. *ACM Transactions on Graphics*, 29(4):1–10, 2010. 3

[34] J. Vandersanden, S. Holl, X. Huang, and G. Singh. Edge-preserving noise for diffusion models. *arXiv*, 2024. 3

[35] A. Voynov, K. Aberman, and D. Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH*, 2023. 3

[36] A. Voynov, Q. Chu, D. Cohen-Or, and K. Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 3

[37] H. Wang, Y. Yu, T. Luo, H. Fan, and L. Zhang. Magic: Multi-modality guided image completion. 2024. 6

[38] S. Wang, C. Saharia, C. Montgomery, J. Pont-Tuset, S. Noy, S. Pellegrini, Y. Onoe, S. Laszlo, D. J. Fleet, R. Soricut, J. Baldridge, M. Norouzi, P. Anderson, and W. Chan. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18359–18369, 2023. 5

[39] Y. Wei, Y. Zhang, Z. Ji, J. Bai, L. Zhang, and W. Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *IEEE/CVF International Conference on Computer Vision*, pages 15897–15907, 2023. 2, 3

[40] G. Xiao, T. Yin, W. T. Freeman, F. Durand, and S. Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 3

[41] S. Xie, Z. Zhang, Z. Lin, T. Hinz, and K. Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023. 3, 5

[42] B. Xue, S. Ran, Q. Chen, R. Jia, B. Zhao, and X. Tang. Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In *European Conference on Computer Vision*, pages 300–316, 2022. 3

[43] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, and F. Wen. Paint by example: Exemplar-based image editing with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 3

[44] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3

[45] R. Yi, H. Tian, Z. Gu, Y.-K. Lai, and P. L. Rosin. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22388–22397, 2023. 6, 7

[46] T. Yu, R. Feng, R. Feng, J. Liu, X. Jin, W. Zeng, and Z. Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 3

[47] B. Zhang, Y. Duan, J. Lan, Y. Hong, H. Zhu, W. Wang, and L. Niu. Controlcom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040*, 2023. 2, 3

[48] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023. 2