DiffusionMix: Adaptive Image Composition through Merging Two DDPM Denoising Processes

Evelyn Kim[1] and Seyun Kim*[1]

[1]CONNECTEVE Inc.

April 3, 2026

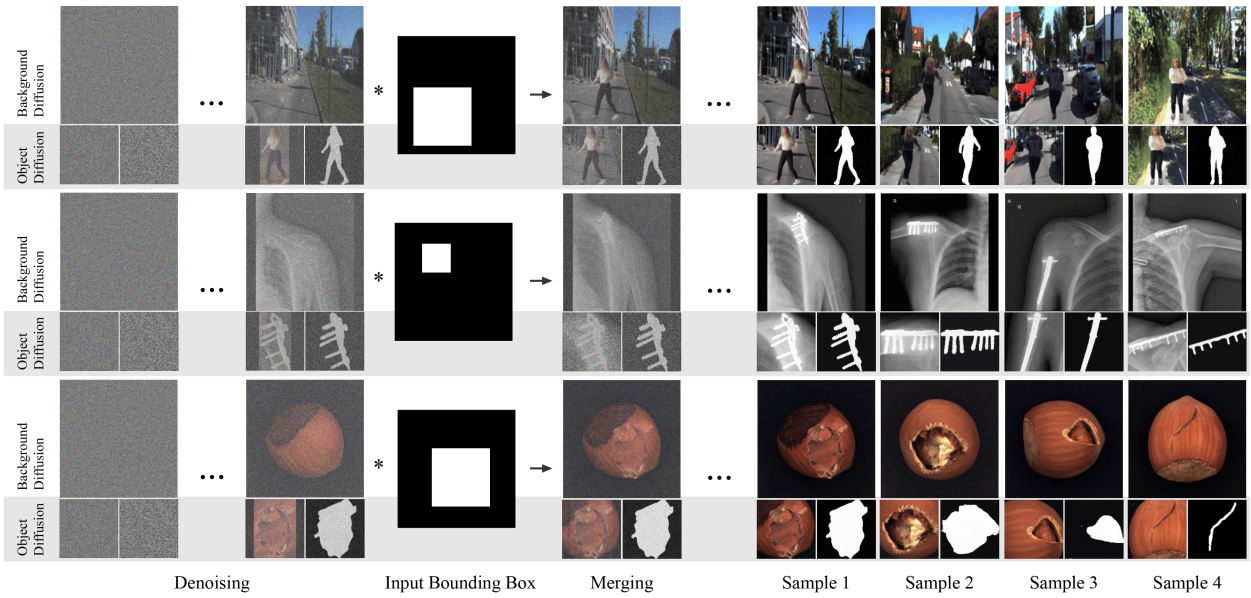


Figure 1: In **DiffusionMix**, two **Denoising Diffusion Probabilistic Models (DDPMs)** are used, one of which is trained with 4-channel (**RGB+Mask**). The **object diffusion model** generates an object image along with a segmentation mask within a given bounding box, while the **background diffusion model** synthesizes the rest of the scene. The two outputs are progressively merged during the denoising process, allowing mutual interaction between object and background, thereby achieving a naturally blended composition.

Abstract

Recent advances in image generation using diffusion models have created highly realistic images. However, most research has focused on generating synthetic images with a single diffusion model, whereas leveraging multiple models offers greater flexibility—allowing modular combinations of specialized models for specific applications and enabling more controlled and diverse image generation. In this work, we propose *DiffusionMix*, a unified framework designed to combine the outputs of two diffusion models. One is trained to generate backgrounds, while the other produces objects and their corresponding masks as 4-channel output (RGB+Mask). This additional channel (mask) enables the generation of detailed segmentation masks, serving as a guide for blending the outputs of both models. To enhance realism, we propose alternating denoising and resampling during the denoising process, where the two models interact and blend their outputs more naturally. Furthermore, modularizing image components allows the retraining of specific parts instead of entire large models, improving computational efficiency and adaptability to various use cases. We demonstrate that DiffusionMix generates high-quality, diverse images while merging outputs and producing segmentation masks. This capability facilitates realistic synthetic datasets for object detection, segmentation, and anomaly detection, while also supporting task-specific image synthesis without requiring

large-scale retraining, enhancing adaptability to a wide range of downstream tasks.

1 Introduction

Diffusion models have emerged as a powerful technology, demonstrating remarkable capabilities in generating high-quality, diverse images. These models have quickly established themselves as the state-of-the-art in image synthesis Saharia et al. [2022b], Croitoru et al. [2022], Rombach et al. [2022b], Song et al. [2023], Dhariwal et al. [2023], Ramesh et al. [2023], dee [2024]. While such advancements hold significant potential for transforming digital content creation, their practical deployment remains a challenge. This is primarily due to the difficulty in customizing these models, which heavily rely on large-scale pre-trained datasets. While such pre-trained models generate highly realistic images, their application to real-world problems often requires more than realism—it necessitates generating task-specific or purpose-driven content. For instance, as illustrated in Figure 5, even when the generated images are visually compelling, they may fail to align with the specific objectives of the target application.

In addition, combining the outputs of two diffusion models is a critical approach for addressing real-world challenges in image synthesis. This method allows for precise control over the generation process, enabling the integration of specific features or objects into desired backgrounds. Such capability is especially valuable when task requirements go beyond generating realistic images to producing purpose-driven content. By leveraging the strengths of individual models—such as one trained on object details and another on backgrounds—this approach ensures that the resulting images align closely with the intended objectives. Furthermore, it offers computational efficiency by avoiding extensive re-training of large pre-trained models, promoting flexibility in adapting to various use cases. This integration enhances creative potential, allowing for the synthesis of diverse styles and contexts that would be challenging for a single model.

We propose DiffusionMix, an adaptive image synthesis framework that leverages two denoising Diffusion Probabilistic Models (DDPMs) to blend their outputs. Instead of training on the standard 3-channel input (RGB), one of the diffusion models is trained with a 4-channel structure (RGB + Mask). This additional channel enables the generation of a segmentation mask, which serves as a guide for blending the outputs of both models. Given a bounding box, the object diffusion model generates an object image along with its corresponding segmentation mask within the specified region, while the background diffusion model synthesizes the remaining part. By combining the bounding box and the segmentation mask, the two outputs are iteratively merged during the diffusion process, allowing the object and background images to influence each other. This interaction ensures a smoother and more natural fusion of the synthesized components.

DiffusionMix, which synthesizes objects and backgrounds has vast potential for diverse applications. In Section 4, we demonstrate the versatility of our approach through examples, including the generation of medical datasets with specific features and the creation of datasets for anomaly detection. These examples show how our framework can be adapted to address real-world challenges, emphasizing its practical utility across various domains such as feature manipulation, background separation and editing, object generation, and data augmentation.

2 Related Works

Diffusion Models Diffusion models are a type of generative probabilistic model that approximate complex data distributions by progressively denoising Gaussian noise. Starting from a noise input $I_T \sim N(0, I)$, these models apply a series of denoising steps to gradually refine the input, transforming it into a sample that follows the target distribution q. Recent studies in diffusion models have pushed the boundaries of generative AI, leading to significant breakthroughs across various domains. The following are representative research areas related to diffusion models.

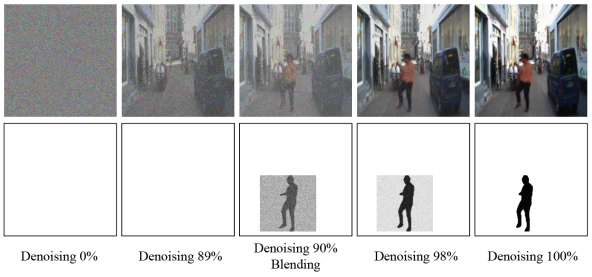
Image-to-Image Translation Diffusion models can convert images from one domain to another while preserving structural information. Denoising diffusion probabilistic models (DDPMs) Ho et al. [2020] laid the foundation for diffusion-based image generation. Building on this, conditional diffusion models have been developed to translate images by conditioning the diffusion process on an input image. Palette Saharia et al. [2022a] is a unified framework for diverse image-to-image translation tasks using conditional diffusion models, achieving state-of-the-art performance in applications like colorization and inpainting. Additionally, Tumanyan et al. [2023] proposes a text-driven image-to-image translation framework that leverages pre-trained diffusion features, generating images from textual descriptions. More recently, Diffi2i Xia et al. [2024] is designed with a novel architecture that reduces the computational cost and training time while maintaining high-quality image generation across various translation tasks.

Text-to-Image Synthesis One of the most actively researched areas in diffusion model studies. Models pre-trained on large datasets are available, and the image generation quality is highly realistic. Along with DDPM Ho

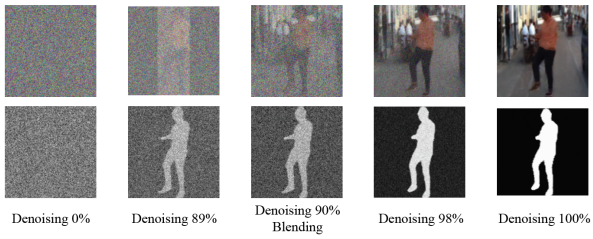
et al. [2020], DALL·E 2 Ramesh et al. [2021] and Stable Diffusion Rombach et al. [2022a] are notable examples, which have demonstrated significant advancements in generating high-quality images from text prompts, achieving impressive results in terms of both fidelity and coherence. DDPM was the foundational model, introducing the denoising diffusion process, which effectively models data distribution by reversing a gradual noising process. DALL·E 2 extended this concept by combining powerful language models with diffusion-based image generation, enabling the creation of detailed images from natural language descriptions. Stable Diffusion performs the diffusion process in a latent space, significantly reducing computational cost while still generating high-quality images. Guided Diffusion Dhariwal and Nichol [2021] further enhanced the flexibility of these models by enabling the incorporation of specific guidance signals, allowing for more controlled and targeted generation.

Super-Resolution and Inpainting Diffusion models have become central to pushing the boundaries of super-resolution Moser et al. [2024]. SR3 Saharia et al. [2022c] performs super-resolution based on iterative refinement using DDPM and achieves competitive results compared to state-of-the-art GAN methods. Srdiff Li et al. [2022] employs a diffusion-based model to gradually transform a Gaussian noise distribution into a high-resolution image through a learned Markov chain process. Similarly, inpainting refers to filling in missing or corrupted parts of an image. RePaint Lugmayr et al. [2022] is a diffusion model-based method for image inpainting while preserving the coherence of the surrounding content. The authors propose a resampling strategy in reverse diffusion for harmonizing the inpainted region with the known region in a single step. More recently, Smartbrush Xie et al. [2023] combines text descriptions and shape guidance to reconstruct missing objects in images. Renderdiffusion Anciukevičius et al. [2023] integrates 3D reconstruction and filling missing regions by applying diffusion processes to both 2D images and 3D data.

Medical Image Analysis Diffusion models are highly effective in medical imaging tasks due to their ability to generate high-quality and detailed samples, making them useful for enhancing diagnostic accuracy, image resolution, and other essential applications Kazerouni et al. [2022, 2023]. These models demonstrate significant potential in both traditional imaging (MRI and CT scans) and more advanced methods, addressing challenges such as noise reduction, image reconstruction, and segmentation in complex medical datasets. The application of diffusion models for medical anomaly detection is explored in Wolleb et al. [2022], with a focus on identifying abnormalities in MRI and X-ray images. The study highlights how Gaussian and



(a) Background Diffusion



(b) Object Diffusion

Figure 2: Denoising process of diffusion models. The bottom row of (b) shows the segmentation mask generated by the object diffusion model, while the bottom row of (a) shows the image with the mask inverted and inserted into the input bounding box. In each case, the outputs of each model fill the white areas.

Simplex noise patterns improve the model's robustness and accuracy. Additionally, diffusion models are gaining popularity in medical image segmentation due to their effectiveness in handling ambiguous segmentation tasks. For instance, MedSegDiff Wu et al. [2024] enhances segmentation accuracy using Dynamic Conditional Encoding and the Feature Frequency Parser (FF-Parser), particularly for low-contrast or ambiguous regions. Similarly, CIMD Rahman et al. [2023] generates multiple segmentation masks from a single input image by incorporating stochasticity at each level within its hierarchical structure.

However, while blending outputs from two diffusion models is an important and promising direction, it remains unexplored. To combine these outputs effectively, further research is needed to train diffusion models with structures beyond the standard RGB channels, such as using 4-channel configurations or other innovative approaches. This remains an open area for exploration, offering opportunities to improve the flexibility and applicability of diffusion models in specialized use cases.

3 Method

Preliminaries

Denoising Diffusion Probabilistic Models (DDPM) Ho et al. [2020] is a generative model that learns to generate data through gradual noise addition and subsequent denoising. This generative approach is based on the idea of reversing a diffusion process, where noise is added incrementally to a data point and the model learns to reverse this noise to recover the original data distribution.

In DDPM, the forward process progressively adds Gaussian noise to an image x_0 over a series of T time steps, resulting in a sequence of increasingly noisy versions of the image, denoted as $x_1, x_2, \ldots, x_T$. The forward diffusion process can be formulated as a Markov chain with the following conditional distributions:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

where β_t is a variance schedule that controls the noise level at each time step t, and $\mathcal{N}$ represents a Gaussian distribution. The process begins with the original data x_0, and iteratively adds noise until the data is fully corrupted into a standard Gaussian distribution at x_T:

$$q(x_T|x_0) = \mathcal{N}(x_T; 0, I) \tag{2}$$

Here, x_T is a sample from a standard normal distribution. Since the noise at each step (1) has independence, the cumulative noise variance is $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. We can thus represent $q(x_t|x_0)$ as follows:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{3}$$

DDPM learns the reverse diffusion process to reverse the noising process and recover the original data. The reverse process is modeled as another Markov chain, parameterized by a neural network. The reverse transition is given by:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I) \tag{4}$$

where $\mu_\theta(x_t, t)$ is the mean predicted by the model, and σ_t^2 represents the variance at each step. The model learns to predict the mean and variance of the reverse process, effectively denoising the corrupted image step by step. To learn this reverse process, the model is trained to minimize the difference between the true posterior distribution $q(x_{t-1}|x_t, x_0)$ and the model's prediction.

This is achieved by optimizing a loss function based on the Kullback-Leibler divergence between the true distribution and the model's prediction over all time steps Ho et al. [2020]:

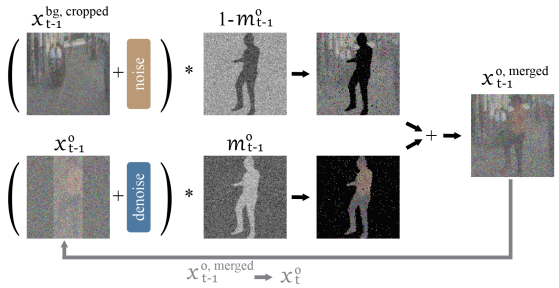


Figure 3: Merging process for object diffusion. The top row shows cropping the background diffusion output, adding noise, and multiplying it by the inverted mask. The bottom row depicts denoising the object diffusion model and applying the mask. The results are then merged as the object diffusion output for the next step. The background diffusion model follows the same process using the masks in Figure 2a.

$$L_T = D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \| p(\mathbf{x}_T)) \tag{5}$$

$$L_{t-1} = D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) \tag{6}$$

$$L_0 = \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \tag{7}$$

$$L = \mathbb{E}_q \left[L_T + \sum_{t>1} L_{t-1} - L_0 \right] \tag{8}$$

The objective function (8) ensures that the model learns to predict the reverse process in a way that accurately recovers the original data from the noisy version.

Once the reverse process is learned, sampling from the model involves initializing with pure Gaussian noise ($x_T \sim \mathcal{N}(0, I)$) and running the reverse diffusion process to generate a sample from the data distribution. This sampling process is typically performed by iteratively applying the learned reverse transitions:

$$x_{t-1} = \mu_\theta(x_t, t) + \sigma_t \epsilon_t$$

where ϵ_t is a noise term sampled from a standard normal distribution. By repeating this process over T steps, the model generates a high-quality sample that is consistent with the original data distribution.

Preparation for the 4-Channel Trained DDPM

In our proposed framework, we use two diffusion models: one for background generation and the other for object creation. For convenience, we will refer to the diffusion model trained on background data as the "background diffusion model", and the one trained on objects or features as the "object diffusion model". Our approach utilizes the DDPM model, without using prompts.

For the background diffusion model, the training process follows the standard approach of learning from RGB images. On the other hand, the object diffusion model is trained on a 4-channel setup, which includes the RGB channels along with an additional segmentation mask channel. The images in the upper row of Figure 2b represent the RGB images generated by the object diffusion model, while the masks in the lower row correspond to the fourth channel, representing the masks generated by the object diffusion model. Both the images and their corresponding object segmentation masks are required to train the object diffusion model. Training with this 4-channel configuration enables the object diffusion model to generate both images and their associated masks simultaneously. Similar to the RGB channels, the model predicts the distribution in the mask channel.

Merging Outputs

We aim to combine the outputs of the object diffusion model and the background diffusion model, performing denoising while reflecting each other to achieve natural image synthesis. The reason for merging the outputs during the denoising process is that simply combining the segmentation masks after denoising results in an unnatural synthesis. When the object and background are combined using masks without any interaction or mutual referencing, the resulting composition lacks cohesion, as if they were simply cut and pasted together. Instead, the background and object should interact with each other, with the background referencing the object and the object being influenced by the background during the denoising process. This interaction helps produce a more natural composition, ensuring the final result appears cohesive and realistic.

Therefore, we merge the images during the denoising process. When combining the images, we use the mask generated by the object diffusion model for the merging process. In the example shown in Figure 2, each model generates its own image up to 89% denoising. The first lines of the figures represent RGB images generated, while the second ones represent masks applied during the merging process. Each model fills in the white areas of the mask, similar to repainting techniques. The masks shown in Figure 2b, m_t^o, are those generated by the object diffusion model, while the masks in Figure 2a, m_t^{bg}, are defined as follows: Up to 89% denoising, there is no mask, as the model generates RGB images without merging two images, thus no mask is needed. When the images are merged, we use the mask defined by adding $1 - m_t^o$ to the bounding box region that we provided as input.

For example, at 90% denoising in Figure 2, a mask produced by the object diffusion model $m_{90\%}^o$ determines how the two images are merged. Let's consider the case where the background portion of the output from the background diffusion model, $x_{90\%}^{bg}$, is inserted into the output of the object diffusion model, $x_{90\%}^o$. $x_{90\%}^{bg}$ is cropped by the input bounding box, as shown in $x_{t-1}^{bg,cropped}$ in Figure 3. Then, this cropped $x_{90\%}^{bg}$ fills the part of $1 - m_{90\%}^o$, and $x_{90\%}^o$ fills the part of $m_{90\%}^o$.

More specifically, the forward process is modeled as a Markov chain that cumulatively adds Gaussian noise (1). Therefore, using (3), we can sample x_t at any time step t. By treating the output of another diffusion model as x_0 and adding noise corresponding to time step t, it becomes possible to merge outputs from a different diffusion model using a mask. Thus, for a single diffusion model, we use (4) for the part that this model is responsible for generating, and we use (3) for the part that corresponds to the output of another model. Thus, during the reverse step of the object diffusion model, the process described in Figure 3 can be expressed as follows:

$$x_{t-1}^{bg,cropped} \sim \mathcal{N}(\sqrt{\bar{\alpha}_t} x_t^{bg}, (1 - \bar{\alpha}_t) I) \tag{9}$$

$$x_{t-1}^{o} \sim \mathcal{N}(\mu_\theta(x_t^o, t), \Sigma_\theta(x_t^o, t)) \tag{10}$$

$$x_{t-1}^{o,merged} = m_{t-1}^o \odot x_{t-1}^o + (1 - m_{t-1}^o) \odot x_{t-1}^{bg,cropped} \tag{11}$$

$x_{t-1}^{bg,\text{cropped}}$ is sampled using the background diffusion model's output x_t^{bg} at step t and cropped to the region defined by the input bounding box. Different masks are applied for the background diffusion case, shown in the second row of Figure 2a.

Resampling

When merging the generated data, the boundaries between elements remain unnatural, highlighting the need for further refinement to address this issue. This issue arises because the regions corresponding to the background and object are sampled independently without considering each other before the image composition. To preserve the object's details, it is preferable to perform image merging at later stages of the diffusion process, once the mask has been sufficiently refined. However, the model's ability to harmonize the content diminishes in later reverse steps, resulting in limited natural blending at the boundaries. This limitation is caused by insufficient interaction between the background and object to create a well-harmonized composition.

To achieve sufficient coherence between the two image components, we utilize a resampling approach that leverages the inherent properties of DDPM to produce consistent structures Lugmayr et al. [2022]. By diffusing the intermediate output x_{t-1} back to x_t, the generated information is exchanged more effectively between the object and background. This operation helps achieve better harmony, particularly at the boundaries

between the two components. To control the resampling process more effectively, we use the concept of a jump length τ. For example, with a total time step of $T = 250$ and $\tau = 10$, the time step sequence is $(250, 249, \ldots, 2, 1, 0, 1, 2, \ldots, 8, 9, 10, 9, 8, \ldots, 2, 1, 0)$, which differs from the standard time step sequence $(250, 249, \ldots, 2, 1, 0)$. We also introduce the concept of the resampling number r, which indicates how many times the process will repeat returning to τ. For example, with a total time step of $T = 250$, $\tau = 10$, and $r = 2$, the time step sequence becomes $(250, 249, \ldots, 1, 0, 1, \ldots, 10, \ldots, 0, \ldots, 10, \ldots, 0)$.

As shown in Figure **??**, longer jump lengths τ and larger resampling number r mean sharing greater information between the background and the object. However, excessive information exchange can negatively impact the generation of high-quality images. For instance, when the object diffusion model produces intricate details, excessive resampling may lead to a loss of precision in the object's representation. To prevent this, tuning of hyperparameters is important to manage a balance between harmonization and preserving fidelity. Thus, we analyze the impact of hyperparameters on the output in Section **??**.

4 Experiments

Implementation details

We trained the object diffusion model and the background diffusion model on different datasets and combined their respective contents. The object diffusion model was trained on the TikTok dataset, our private shoulder X-ray dataset, and anomaly detection datasets, including MVTec Bergmann et al. [2019] and VisA Zou et al. [2022]. The background diffusion model was trained on the KITTI, ImageNet, and COCO datasets, our X-ray dataset, MVTec, and Visa.

We used 256×256 image size for the background model, while the object diffusion model was trained with images of 64×64 or 128×128 sizes depending on the dataset. For ImageNet, we used a pre-trained guided diffusion model, while we trained directly for the other datasets on a single NVIDIA H100 GPU.

Empirical Analysis on Multiple Datasets

The results of applying our method to diverse datasets are demonstrated in Figure 1. The first row illustrates the composition of two diffusion models, the background model trained on the KITTI dataset and the object model trained on the TikTok dataset. The second row shows a case where the background diffusion model was trained on shoulder X-ray images without implants, and the object diffusion model was trained on implant objects. The third row uses the hazelnut subset of the MVTec dataset, where the background diffusion model was trained on normal hazelnut images, and the object diffusion model was trained on anomalies. Additionally, further results combining the KITTI and TikTok datasets, as well as integrating TikTok with the background diffusion model trained on COCO and ImageNet, are presented in Figure 4. Notably, the diversity of both the objects and backgrounds is high, with masks enabling blends that preserve object details while maintaining a natural appearance.

Qualitative Comparison for Practical Tasks

We compared several models capable of generating object-specific content based on a given bounding box with our DiffusionMix, as shown in Figure 5. The models include MultiDiffusion Bar-Tal et al. [2023], InstantDiffusion Wang et al. [2024], and HiCo Cheng et al. [2024]. These models rely on large-scale pre-trained networks and tend to generate abstract images. While they can produce diverse outputs based on prompts, they have limitations in generating realistic images tailored to specific services. In contrast, our model produces highly realistic images, which is an expected outcome since it is exclusively trained on actual medical data, unlike other models that lack such specialized training. However, fine-tuning these other models using real medical images presents significant challenges. Most of these models require additional structured datasets containing prompts or other supplementary information for training or fine-tuning, which complicates the process. Our approach, on the other hand, uses a simple DDPM without prompts, making training straightforward and efficient. This allows easy training with just image data, without carefully designed prompts or complex structures. As a result, our model produces highly realistic images suited for practical applications, such as creating datasets to train AI models for implant detection, segmentation, and classification in real-world scenarios.

Application on Anomaly Detection Datasets

One of the key applications of DiffusionMix is generating anomaly detection datasets. By using an object diffusion model for anomalies and a background diffusion model for normal images, their outputs combine to produce anomaly images with corresponding segmentation masks. This method goes beyond simple overlaying, allowing anomalies and backgrounds to interact during generation, resulting in more realistic outputs. Figure 6 shows anomaly images generated with MVTec Bergmann et al. [2019]. During denoising, the object and background influence each other, gradually adjusting to reflect mutual changes. For example, in the Hazelnut dataset, anomalies blend into normal hazelnut images while the background

Figure 4: Composition of various background diffusion models (256x256) and an object diffusion model trained on the TikTok dataset.

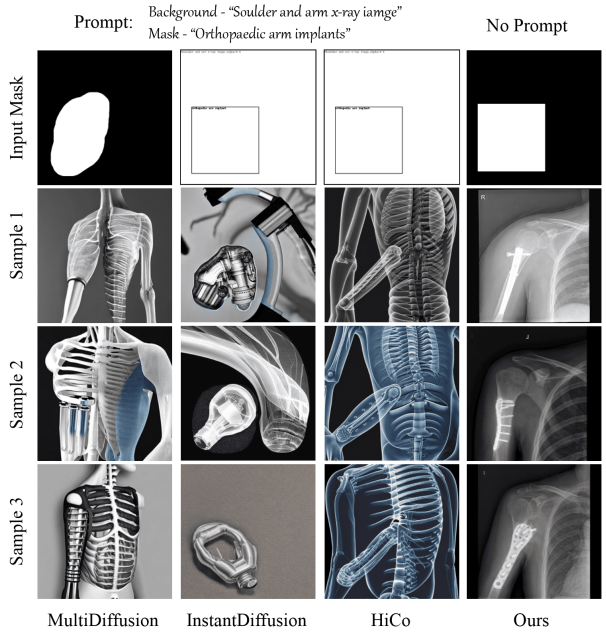


Figure 5: Qualitative comparison of DiffusionMix and other methods.

adapts to accommodate them. Notably, in Crack anomalies (e.g., Sample 3 and Sample 4), even when the anomaly deviated from the hazelnut's round shape, the hazelnut wrapped around it, producing realistic outputs. Similarly, in the Carpet dataset, woven patterns naturally connected between the anomaly and background, demonstrating organic integration. This was achieved through resampling, where anomalies and backgrounds were generated interactively. For Carpet examples, the object diffusion model was trained on 128-sized images for hole and metal anomalies and 64-sized images for cut and color anomalies. Additionally, Figure 7 shows examples of anomaly data generated using the VisA Zou et al. [2022]. The resampling process allowed for interactive generation between the background and the anomalies, producing realistic outputs. This capability to generate a large variety of anomaly images holds great potential for creating datasets that can be effectively used for anomaly detection tasks.

5 Conclusion

In this paper, we introduced the DiffusionMix framework, which combines and harmonizes the outputs of two DDPM models during the denoising process. By designating the models as an object diffusion model and a background diffusion model, our approach focuses on naturally integrating objects into backgrounds. The results highlight the effectiveness of our method in generating realistic and cohesive composite images by leveraging both models. Additionally, a resampling mechanism allows us to control the degree of interaction between the object and background, enabling more adaptable and context-aware image generation. With these capabilities, DiffusionMix opens up new possibilities for tasks that demand realistic object-background integration.

Limitation Our model is based on DDPM, which is inherently much slower compared to GAN-based or autoregressive methods, making it less suitable for real-time applications. However, as DDPMs are becoming more popular, recent research has been actively addressing efficiency improvements, suggesting that this limitation may gradually be mitigated over time. Secondly, unlike the current trend, our method does not rely on prompts with large pre-trained models for scene generation. As such, further research is needed to explore how prompts can be integrated into this framework to expand its capabilities and align with recent advancements.

Social impact aspect DiffusionMix is designed to composite contents from different models. However, it is important to acknowledge the potential for misuse, such as

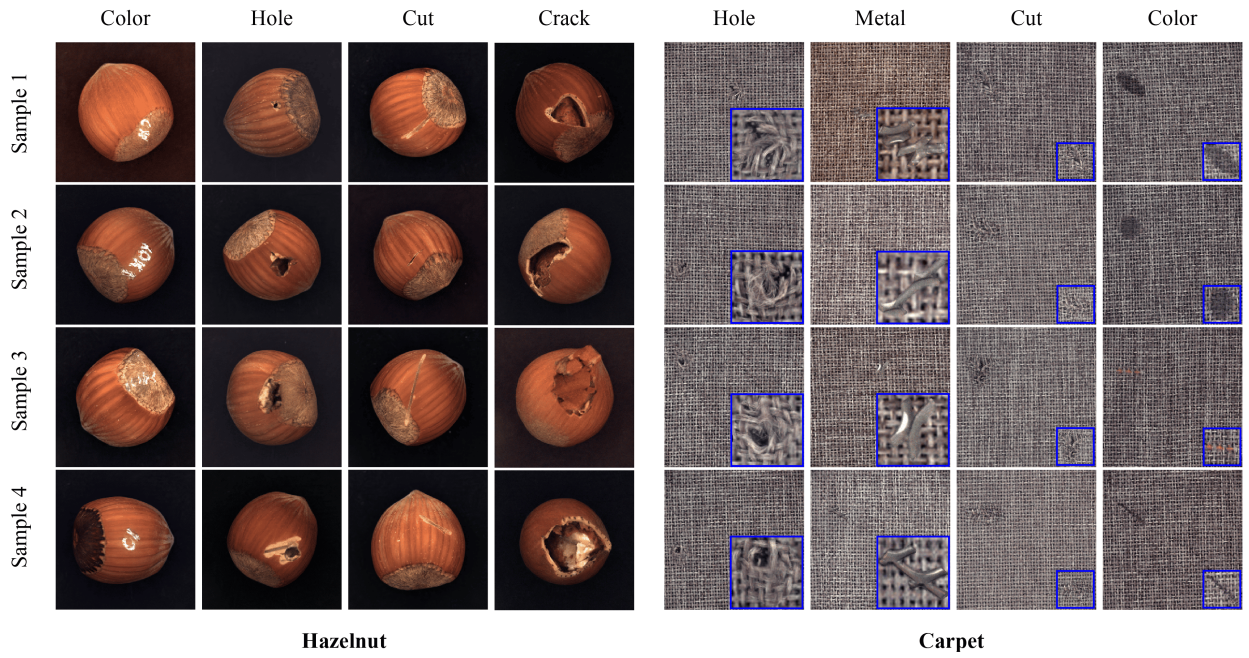


Figure 6: Visual results for the composition of anomaly (object diffusion) and normal image (background diffusion) on MVTec. Background models with 256-size images and object diffusion models with 128-size and 64-size images were used.

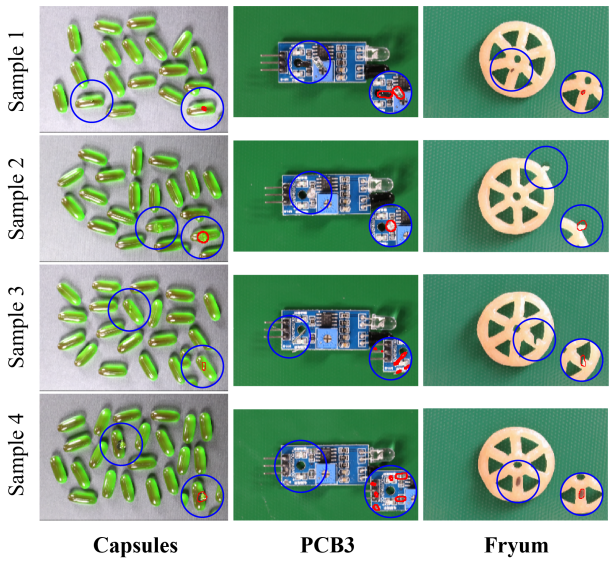


Figure 7: Visual results for the composition of anomaly (object diffusion) and normal image (background diffusion) on VisA. The image within the circle represents the segmentation results.

combining inappropriate or sensitive content. To address this concern, we emphasize the importance of regulating the use of such models and developing tools for detecting misuse. Taking these measures will support the responsible advancement of AI technology for the betterment of humanity.

References

Deep diffusion models for image synthesis. *arXiv:2403.12743*, 2024. URL `https://arxiv.org/abs/2403.12743`.

Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12608–12618, 2023.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

Bo Cheng, Yuhang Ma, Liebucha Wu, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. Hico:

Hierarchical controllable diffusion model for layout-to-image generation. *arXiv preprint arXiv:2410.14324*, 2024.

Ionel Croitoru, Marius Leordeanu, Nasir Rahaman, et al. Diffusion models in vision: A survey. *arXiv preprint arXiv:2209.04747*, 2022.

P. Dhariwal, A. Nichol, et al. Efficient diffusion models for fast image generation. *arXiv preprint arXiv:2301.05493*, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

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

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models for medical image analysis: A comprehensive survey. *arXiv preprint arXiv:2211.07804*, 2022.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023.

Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

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

Brian B Moser, Arundhati S Shanbhag, Federico Raue, Stanislav Frolov, Sebastian Palacio, and Andreas Dengel. Diffusion models, image super-resolution, and everything: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11536–11546, 2023.

A. Ramesh, P. Dhariwal, et al. Conditional diffusion models for text-to-image synthesis. *arXiv preprint arXiv:2304.12494*, 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

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, 2022a.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, et al. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2022b.

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

Chitwan Saharia, William Chan, Saurabh Saxena, et al. Imagen: Text-to-image diffusion models with unprecedented photorealism. *arXiv preprint arXiv:2205.11487*, 2022b.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45 (4):4713–4726, 2022c.

Y. Song, J. Sohl-Dickstein, et al. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2304.03275*, 2023.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6232–6242, 2024.

Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *International Conference on Medical image computing and computer-assisted intervention*, pages 35–45. Springer, 2022.

Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic

model. In *Medical Imaging with Deep Learning*, pages 1623–1639. PMLR, 2024.

Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, Radu Timotfe, and Luc Van Gool. Diffi2i: Efficient diffusion model for image-to-image translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

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

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022.