Scene Generation for Object Rearrangement with Latent Diffusion Models

Hogun Kee and Songhwai Oh

Department of Electrical and Computer Engineering and ASRI, Seoul National University,
Seoul, 08826, Korea (hogun.kee@rllab.snu.ac.kr, songhwai@snu.ac.kr)

Abstract: In this paper, we address goal scene generation for the object rearrangement problem. We introduce a novel scene generation method leveraging latent diffusion models (LDM) to create goal images that guide the rearrangement process. Our approach enables the generation of goal configurations while preserving the inherent characteristics of objects, such as shape, size, and texture, overcoming common limitations in traditional image generation methods. Specifically, we utilize a variational autoencoder (VAE) to extract relevant object features from images, which are then used to condition the LDM, ensuring that the newly generated configurations align with the physical realities of the existing objects. We demonstrate the effectiveness of our method through various qualitative results, showing that our system can generate goal images for object rearrangement while maintaining object consistency.

Keywords: Artificial Intelligent Systems, Robot Vision, Robotic Applications

1. INTRODUCTION

In robotic manipulation, object rearrangement is a significant challenge. The problem is defined as moving objects to meet specific conditions. Depending on the conditions provided, this can be defined as a sorting problem, where similar types of objects are grouped together, or as a stacking problem, where objects are layered on top of each other. When a target arrangement is directly provided, it becomes a goal-conditioned object rearrangement problem. For such cases where the goal is explicitly provided, numerous studies have been conducted on problems where the goal is given in the form of an image or where specific target positions for each object are specified.

When the task goals are provided in image form, each task has a clear goal, which has led to extensive research in goal-conditioned reinforcement learning and task and motion planning (TAMP). However, such problem settings have limitations when applied to the real world, primarily because the goal information must be known in image form, necessitating prior arrangement to set the goal. This means that goals can only be set for arrangements that have already been made, creating a cumbersome and contradictory process for setting new goals, which involves arranging objects as per the new configuration and then capturing the image.

To overcome these limitations, we conducted research to generate goals for object rearrangement. We used latent diffusion models (LDM) [1], a type of image generation method, to create goal images. Initially, we trained a Variational Autoencoder (VAE) [2] to extract object features from images. Subsequently, we used these extracted object features as conditions to train the LDM to generate new configurations reflecting the existing objects' information. Examples of the reconstructed images, maintaining object information, are displayed in Figure 1. Ultimately, the goal of this research is to generate goal images that match new configurations of objects while preserving the information of the current objects on the tabletop.

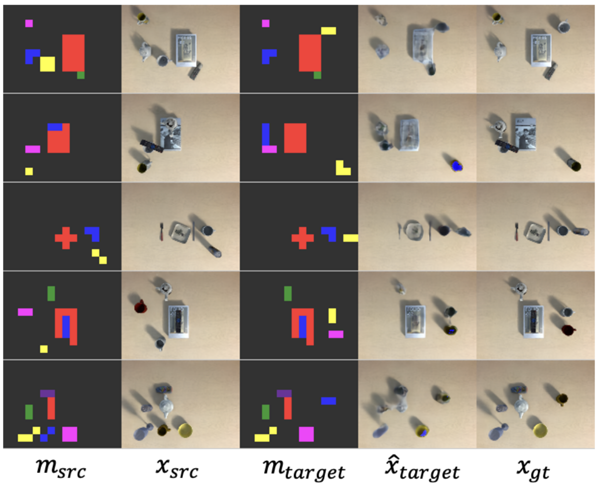


Fig. 1. This is an example of generating a scene image with a new configuration while preserving the characteristics of the objects. m_{src} is the mask of the source image and m_{target} is the target configuration, where the same color means the same object. Our goal is to generate a new scene image according to m_{target} using the visual information from x_{src}. We can compare the generated result with the ground truth scene image x_{gt}.

2. RELATED WORK

Image generation has made significant advances due to research in GAN methods and diffusion models. However, traditional image generation techniques face several challenges when applied to object rearrangement problems due to various reasons.

The first challenge is the need to maintain existing object information while creating a new configuration. Image editing methods like [3] and [4] have enabled the creation of new images by applying changes to current images through language instructions or various forms of conditioning. However, while these methods can replace existing objects with new ones or add objects to empty spaces, they have shown poor performance in tasks that involve moving an existing object to a different location within the image. This problem of object switching has been identified as a limitation in many image generation approaches.

The second challenge is that for the generated results to serve as goals for object rearrangement, it is essential to maintain the size, shape, and textural vision information of the existing objects while only altering their positions and rotations. Traditional image generation methods have demonstrated poor performance in relocating existing objects to different positions while preserving their form and size. In the [5], efforts have been made to overcome these limitations by using object matching with the current object set in the generated images, enabling more robust operation even when objects of different shapes and sizes appear.

The goal of this research is to generate goal images that match new configurations of objects on the current tabletop while preserving the information of the existing objects. To achieve this, we first trained a VAE to extract object features from images. Subsequently, we used these extracted object features as conditions to train a LDM to create new configurations that reflect the existing objects' information.

3. METHODS

3.1 Training Object Feature with Variational Auto-Encoder (VAE)

To train a diffusion model in latent space, we first trained a VAE that transforms RGB images into latent space. We designed a VAE based on ResNet [6] to accept 128x128 RGB images and produce features with 16 channels in a 16x16 format. The features extracted by the VAE contain visual information about the objects placed on the table. These features are used in the LDM during the diffusion process and also serve as the data input for the conditioning vector, which provides information about the objects.

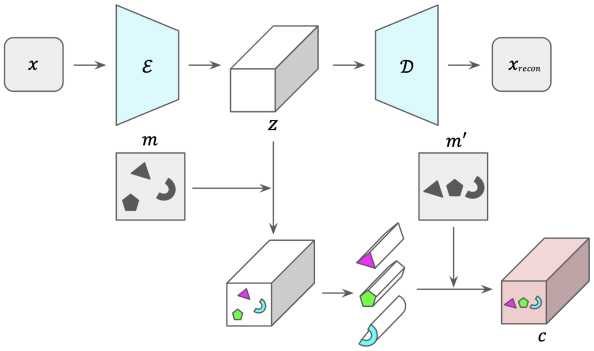


Fig. 2. We assume that the pixels corresponding to each object's segmentation mask on the table contain the core information of that object. The core features are transferred directly to create a new feature map, which is then used as the conditioning vector.

3.2 Conditioning Methods for Object Consistency

To provide object information as a condition for LDM, we utilized object features extracted by a VAE. A pre-trained VAE is used to extract a 16x16 object feature map.

As illustrated in Figure 2, we first extract a latent vector z from the current image x using encoder E. By examining the current table's segmentation mask m, we extract features belonging to each object's mask and identify these as the features for the respective objects. We then create a new latent vector c, which has the same dimensions as z but is initialized with all zeros. The object features are repositioned according to a new mask m' and placed into the new latent vector c. This latent vector c serves as the conditioning vector in the LDM, reflecting the information of the objects.

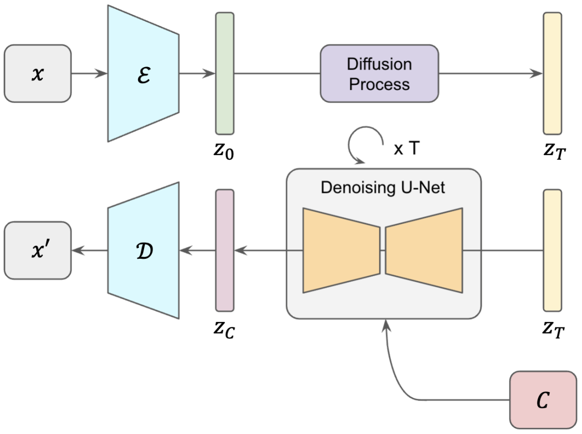
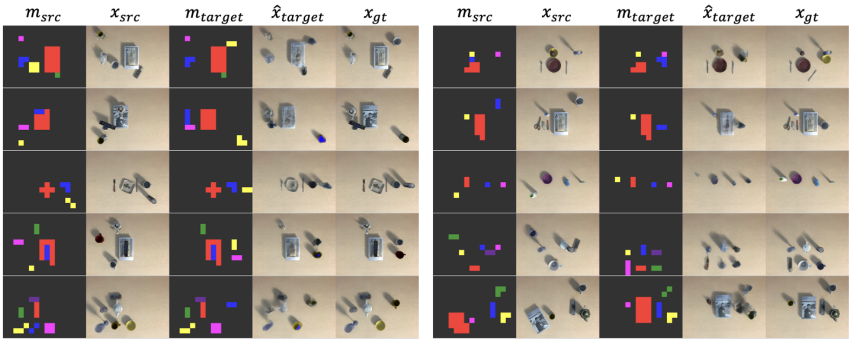


Fig. 3. The diffusion process is conducted in a latent space. In conventional LDM, the conditioning vector is incorporated into the U-Net using a self-attention method. However, we directly concatenated the grid-shaped object feature map into the middle layers of the U-Net to provide conditioning.

3.3 Scene Generation with Latent Diffusion Models

Diffusion models consist of a forward process known as the noising process, where noise is gradually added to clean data, and a reverse process that involves denoising by gradually removing the noise to converge towards the original data distribution $p(x)$, thereby generating new data samples.

The LDM operates the diffusion process in a latent space which can be obtained through the pre-trained VAE. This allows for computation in a much smaller space than the image space, making the process significantly more efficient. We embed a 128x128 RGB image into a 16x16 latent vector, then condition it with the object features c obtained in section 3.2 and go through a 1000-step denoising process to obtain z_c. Decoding this z_c through the trained decoder D results in a scene image with a new configuration while maintaining object consistency. This process is illustrated in Figure 3.



(a) Successful cases

(b) Marginal cases

(c) Failed cases

Fig. 4. Among the results generated through the Latent Diffusion Model (LDM), there are successful cases, failed cases, and marginal cases. Successes were considered when the generated images closely resembled the ground truth images according to the target mask m_{target}. Failures were identified when objects disappeared, the shapes of objects were unclear, or they were not properly regenerated.

4. EXPERIMENTS

4.1 Training VAE with Simulation Data

To train the VAE, we collected tabletop scene images using the PyBullet simulator. We trained the VAE with 128x128 RGB images and obtained the reconstructed image $x_{\text{recon}} = D(E(x_{\text{src}}))$ for the original image x_{src}, where D and E represent the decoder and encoder of the VAE, respectively. The image reconstruction results of the trained VAE are shown in Figure 5. The results demonstrate that the characteristics of the objects are largely preserved in the reconstructed images.

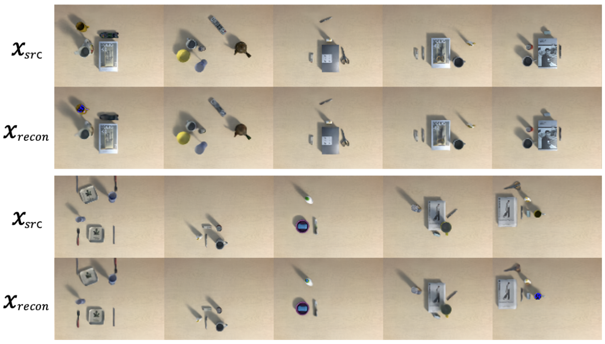


Fig. 5. It can be observed that the reconstruction preserves the size, color, and shape of each object, including the background, effectively.

4.2 Scene Generation with Object Conditioning

The results of scene generation using object features as conditioning are displayed in Figure 4. In successful cases, the original characteristics of the objects are well-maintained, and each object is accurately arranged according to the target mask m_{target}. However, in failed cases, some objects disappear or appear indistinctly shaped during the reconstruction. Additionally, multiple objects may fuse into clumps, creating configurations that are difficult to use for object rearrangement.

In marginal cases, objects are often regenerated as different but category-consistent objects, suggesting that our object conditioning vector does not merely learn visual information but also captures objectness. However, it has been observed that in many cases, either the precise location is not identified, or even if the location is accurate, the distinct characteristics of the objects are not clearly manifested.

5. CONCLUSIONS

In this paper, we propose a method for generating scene images of new configurations by relocating objects while preserving their characteristics. Our approach is based on Latent Diffusion Models (LDM) and uses object features extracted by a Variational Autoencoder (VAE) as conditions. We have confirmed that this method helps in maintaining the visual information of the objects.

However, there have been cases where the precise location of objects was not accurately predicted, and the shape, size, and color information of objects were not correctly reflected, indicating a need for further development in maintaining object consistency.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO. RS-2021-II212068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University)]

REFERENCES

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

[2] Diederik P. Kingma and Max Welling, "Auto-encoding variational Bayes." *arXiv preprint arXiv:1312.6114*, 2013.

[3] T. Brooks, A. Holynski, and A. Efros, “Instruct-pix2pix: Learning to follow image editing instructions.” *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* IEEE, 2023, pp. 18392-18402.

[4] L. Zhang, A. Rao, and M. Agrawala. “Adding conditional control to text-to-image diffusion models.” *in Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2023, pp. 3836-3847.

[5] I. Kapelyukh, V. Vosylius, and E. Johns, “Dall-e-bot: Introducing web-scale diffusion models to robotics.” *IEEE Robotics and Automation Letters.* vol. 8, no. 7, pp. 3956-3963, 2023.

[6] K. He, X. Zhang, S. Ren, S., and J. Sun, “Deep residual learning for image recognition.” *in Proceedings of the IEEE conference on computer vision and pattern recognition.* IEEE, 2016, pp. 770-778.