Replace-in-Ego: Text-Guided Object Replacement in Egocentric Hand-Object Interaction

Minsuh Song*
Sogang University
thdalstj97@sogang.ac.kr

Junho Park*
AI Lab, LG Electronics
junho18.park@gmail.com

Suk-Ju Kang†
Sogang University
sjkang@sogang.ac.kr

Abstract

Editing objects in hand-object interaction scenes in an egocentric view is challenging, as it requires precise localization of the target object while preserving the surrounding context, under frequent hand occlusions and dynamic camera perspectives. Existing inpainting approaches often depend on manual masks or coarse bounding boxes, leading to inaccurate boundaries and visual artifacts. In this work, we introduce a text-guided object replacement framework, Replace-in-Ego, which integrates a vision-language model (VLM)-based segmentation model with a diffusion transformer (DiT). Given target, reference images and the corresponding descriptive texts, our method predicts segmentation masks for the specified objects and applies them to create masked target and reference images. The masked images are merged to encode both target scene context and reference object appearance and served as a joint representation, which conditions DiT to generate a realistic replacement in the target scene. Experiments on TACO dataset demonstrate our approach achieves high-quality reconstruction in egocentric hand–object interaction scenarios, producing sharper boundaries, coherent object placements, and visually realistic integration.

1. Introduction

Accurately editing objects in egocentric hand–object interaction scenes is an essential capability for applications such as augmented and virtual reality, human–robot collaboration, and immersive media content creation. A common requirement in these scenarios is to replace a specific object while keeping the rest of the scene intact. Conventional inpainting methods [2, 9, 12, 19] often depend on manually drawn annotations like bounding boxes or pixel-wise masks, which are labor-intensive and prone to inaccuracies. In particular, approaches that rely solely on bounding boxes frequently produce artifacts and imprecise object boundaries after editing. These challenges are even more pronounced in egocentric views, where frequent hand occlusions and dynamic camera perspectives make accurate replacement highly demanding.

To address these issues, we present a text-guided object replacement framework, *Replace-in-Ego*, which leverages vision–language model (VLM)-based segmentation model [8] in combination with a diffusion transformer (DiT) [14], as shown in Fig. 1. With the recent advancement of VLM [1, 8, 10, 18], it has become possible to obtain segmentation masks directly from natural language prompts, eliminating the need for manual labeling. Moreover, unlike conventional U-Net [16] based diffusion models such as Stable Diffusion [15], our framework adopts DiT to exploit powerful and reliable image generation capability. By representing latent features as patch tokens and modeling their relationships through self-attention, DiT captures global scene context more effectively, leading to improved structural consistency and sharper object boundaries in complex replacement tasks.

Specifically, our *Replace-in-Ego* first takes target, reference images and the corresponding descriptive texts, and predicts segmentation masks for the respective objects. The masked target and reference images are concatenated to serve as a compact representation by encoding both the target scene context and the visual details of the reference object. The concatenated image is processed by a frozen VAE encoder [5], while the reference mask is separately embedded by a lightweight mask encoder. DiT then integrates these embeddings with sampled noise to generate a new target image where the specified object is replaced by the reference object in a seamless manner. To evaluate *Replace-in-Ego*, we adopt TACO dataset [11] due to its diversity of hand-object interactions from a first-person perspective, and showcase *Replace-in-Ego* yields superior reconstructions in egocentric hand-object interaction scenes, delivering cleaner boundaries, consistent object placement, and realistic visual integration.

[*] Equal contribution.
[†] Corresponding author.

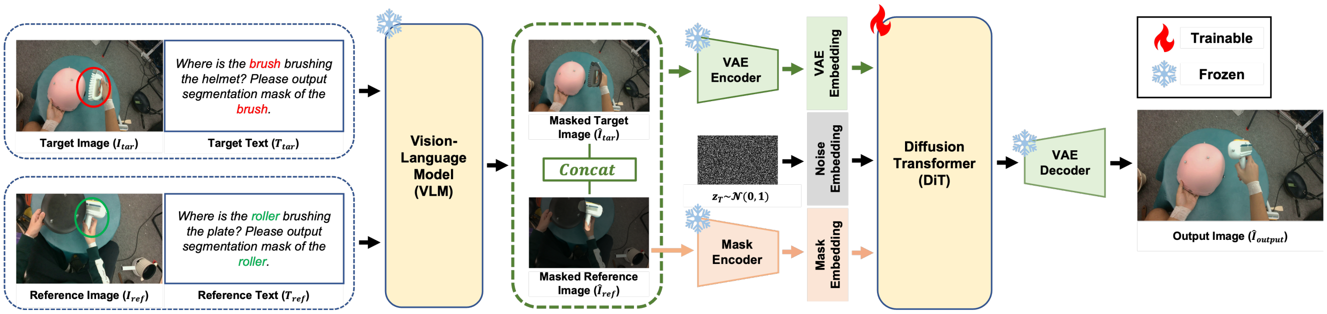
Where is the brush brushing the helmet? Please output segmentation mask of the brush.
Target Image (I_{tar})
Target Text (T_{tar})
Where is the roller brushing the plate? Please output segmentation mask of the roller.
Reference Image (I_{ref})
Reference Text (T_{ref})
Vision-Language Model (VLM)
Masked Target Image ($\hat{I}_{tar}$)
Concat
Masked Reference Image ($\hat{I}_{ref}$)
VAE Encoder
$z_T \sim \mathcal{N}(0,1)$
Mask Encoder
VAE Embedding
Noise Embedding
Mask Embedding
Diffusion Transformer (DiT)
VAE Decoder
Trainable
Frozen
Output Image ($\hat{I}_{output}$)

Figure 1. **Overall pipeline of *Replace-in-Ego*.** *Replace-in-Ego* takes target and reference images (I_{tgt} and I_{ref}) with corresponding texts (T_{tgt} and T_{ref}), produces masked images ($\hat{I}_{tgt}$ and $\hat{I}_{ref}$) using an off-the-shelf vision-language model (VLM) [8], and encodes them via a VAE [5] and mask encoder to obtain a VAE and mask embeddings (z_{vae} and z_{mask}). A diffusion transformer (DiT) [14] then fuses these embeddings with a random noise to generate an output image $\hat{I}_{output}$, which contains target scene with the replacement of the reference object, via a VAE decoder.

2. Method

2.1. Overall Architecture

As illustrated in Fig 1, our pipeline, *Replace-in-Ego*, first takes a target image I_{tgt} and a target text T_{tgt} specifying the object to be replaced. In addition, a reference image I_{ref} and a reference text T_{ref} are also fed as inputs. Next, an off-the-shelf vision–language model (VLM) [8] predicts corresponding binary masks M_{tgt} and M_{ref} for I_{tgt} and I_{ref}, isolating the described objects. M_{tgt} and M_{ref} are used to obtain a masked target image $\hat{I}_{tgt}$ and a masked reference image $\hat{I}_{ref}$, which are concatenated to form a compact representation $\hat{I}_{concat}$ for a diffusion transformer (DiT) [14]. $\hat{I}_{concat}$ is then encoded by a frozen VAE [5] to produce a VAE embedding z_{vae}, which contains a contextual feature. In parallel, $\hat{I}_{ref}$ is fed into a separate mask encoder [20] to extract a mask embedding z_{mask}, which has a shape-specific feature. Finally, DiT fuses z_{vae}, z_{mask}, and a random noise embedding to generate an output image $\hat{I}_{output}$, which contains the target scene and the reference object in place, via a VAE decoder.

2.2. Vision-Language Model

To obtain binary masks of objects (M_{tgt} and M_{ref}) without manual annotation, we employ the pre-trained VLM capable of segmentation from images (I_{tgt} and I_{ref}) and texts (T_{tgt} and T_{ref}). I_{tgt} and I_{ref} are first processed by SAM [6] as a vision backbone which then produces high-quality visual features. These features, along with T_{tgt} and T_{ref}, are fed into a multi-modal LLM [10], which is equipped with a pretrained LoRA [4] module for efficient adaptation to segmentation tasks without modifying the frozen backbone parameters. This architecture allows *Replace-in-Ego* to directly infer segmentation masks from natural language descriptions, enabling flexible object selection without manual annotations.

Following the prediction of binary masks from VLM, we produce $\hat{I}_{tgt}$ and $\hat{I}_{ref}$ by applying M_{tgt} and M_{ref} to I_{tgt} and I_{ref}, respectively. In $\hat{I}_{tgt}$, the pixels corresponding to the object to be replaced are removed, while the background is kept intact to maintain the spatial context of the scene. In $\hat{I}_{ref}$, only the replacement object remains visible, with all other pixels set to zero, isolating the object's visual characteristics. The overall process is formulated as follows:

$$\hat{I}_{tgt} = I_{tgt} \odot (1 - M_{tgt}), \quad \hat{I}_{ref} = I_{ref} \odot M_{ref}, \tag{1}$$

where $\odot$ denotes pixel-wise multiplication.

2.3. Diffusion Transformer

For training and inference for DiT, we first extract the VAE and mask embeddings (z_{vae} and z_{mask}) by the VAE and mask encoder. Specifically, $\hat{I}_{tgt}$ and $\hat{I}_{ref}$ are concatenated along with the width dimension and $\hat{I}_{concat}$ is obtained. This target-reference composite maintains spatial layout information from the target scene while underlying the complete visual cues of the reference object. $\hat{I}_{concat}$ is subsequently passed into the frozen VAE encoder to produce z_{vae}. In parallel to VAE encoding, $\hat{I}_{ref}$ is separately processed by the mask encoder to yield z_{mask}. It provides fine-grained object-centric information to guide the replacement process.

During training, we follow forward and reverse diffusion process as general diffusion models [3, 17]. First, in the forward process, Gaussian noise is incrementally added to the clean latent z_0, which is encoded from a ground-truth image by VAE encoder, over T timesteps. At a randomly sampled timestep t, a noisy latent z_t is obtained as follows:

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{2}$$

where $\bar{\alpha}_t$ denotes a noise level of t.

Next, the reverse diffusion process iteratively generates the output latent, which contains the target scene with the

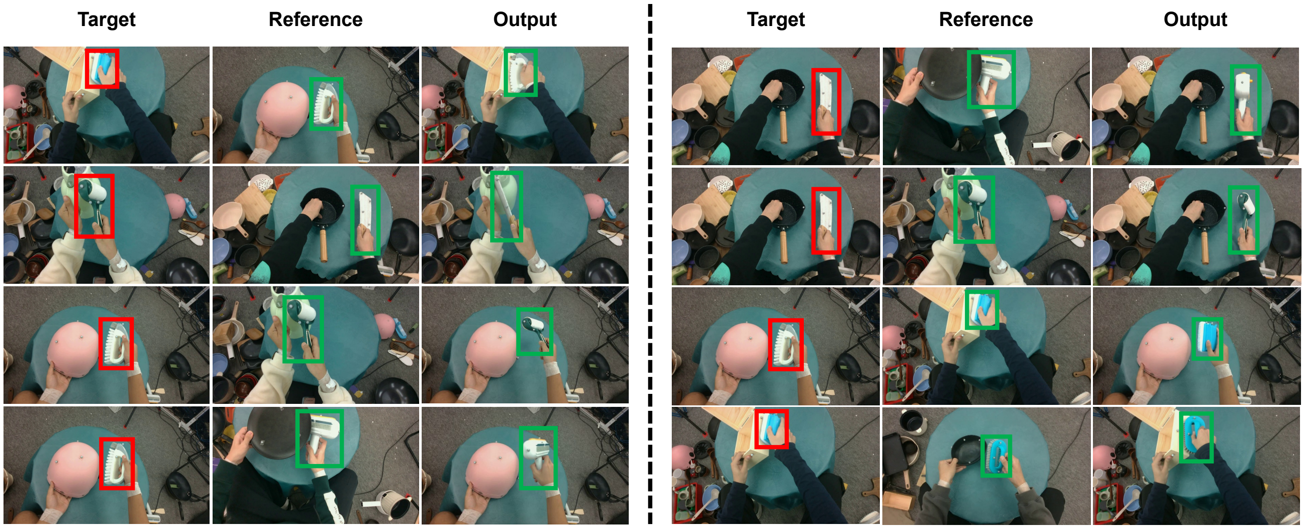

Figure 2. **Qualitative results of *Replace-in-Ego*.** Each row shows two insertion tasks, first and fourth columns indicate the target object (red box), second and fifth columns provide the reference object from another scene (green box), and third and sixth columns display the output generated by *Replace-in-Ego* (green box).

reference object seamlessly integrated. Unlike convolutional U-Net [16] based denoisers [15], DiT processes latent conditions as patch tokens, enabling each transformer block to capture both fine-scale local details and long-range spatial relationships through self-attention and cross-attention mechanisms. Specifically, at each timestep t, DiT receives z_t, which is generated by the forward process, z_{vae}, which contains the context of target scene and reference appearance, and z_{mask}, which is derived from the reference object mask. These conditioning signals are tokenized and embedded into the transformer layers, where cross-attention modules allow the noisy latent tokens to query relevant features from z_{vae} and z_{mask}. Consequently, the forward and reverse processes for a denoising network ϵ_θ are carried out to predict ϵ_t with following objective:

$$\mathcal{L} = \mathbb{E}_{z_0,t,\epsilon_t}\left[\|\epsilon_t - \epsilon_\theta(z_t, z_{\text{vae}}, z_{\text{mask}}, t)\|_2^2\right]. \qquad (3)$$

Minimizing this loss encourages ϵ_θ to accurately predict noise across different timesteps, improving robustness and reconstruction fidelity.

During inference, the pre-trained ϵ_θ progressively denoises a pure Gaussian noise to obtain final latent, and it is decoded to the output image $\hat{I}_{output}$ with the VAE decoder. Therefore, this transformer-based design allows the generator to maintain global scene consistency while integrating the reference object into the target scene with realistic boundaries and textures.

3. Experiments

3.1. Implementation Details

For the dataset, we adopted TACO [11], which provided diverse egocentric recordings with natural hand-object interactions. All images were pre-processed at pixel resolution of 768 × 768. In VLM, a vision backbone was SAM [6], and a language backbone was LLaVA [10] with pretrained LoRA [4]. In DiT, we designed it based on FLUX.1 Fill [dev] [7], a DiT-based generative model. We optimized DiT using Prodigy Optimizer [13] with a weight decay of 0.01, a batch size of 4, and train for 30 epochs on a single NVIDIA H100 GPU. We followed the standard noise prediction loss of DDPM [3] as training objective to ensure stable convergence.

3.2. Results

As shown in Fig. 2, we showed qualitative results of our model, *Replace-in-Ego*. Each row shows two insertion tasks, first and fourth columns indicate target objects, second and fifth columns provide reference objects from another scene, and third and sixth columns display output images generated by *Replace-in-Ego*. The generated results consistently demonstrated *Replace-in-Ego* not only preserves the shape and appearance of the reference object, but also seamlessly adapts to the context of the target scene, even in egocentric perspectives involving complex hand–object interactions. Specifically, *Replace-in-Ego* successfully handled challenging cases involving variations in object scale, orientation, and background clutter. For example, when the reference object was rotated or partially occluded, the generated output still produced a coherent object placement without noticeable artifacts. Moreover, the boundaries between the inserted object and the surrounding scene were well-aligned, indicating *Replace-in-Ego* effectively leveraged both the VAE and mask representations. Therefore, these results confirmed *Replace-in-Ego* achieved realistic and contextually consistent object insertion in first-person interaction scenarios.

4. Conclusion

We present *Replace-in-Ego*, an end-to-end framework for object replacement that integrates segmentation-aware vision-language model (VLM) with a diffusion transformer (DiT). *Replace-in-Ego* targets egocentric hand-object interaction scenarios, which are challenging due to frequent hand occlusions and dynamic camera perspectives. We address this issue by combining multi-modal features obtained from VLM. The combined representation encodes both the target scene context and the reference object appearance, guiding DiT to generate a realistic replacement under diverse egocentric views. Experimental results demonstrate *Replace-in-Ego* effectively preserves scene context while generating semantically faithful output, marking a significant step toward controllable generative modeling for complex everyday interactions.

Acknowledgements

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00414230) and MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2025-RS-2023-00260091) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

[1] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 1

[2] Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. Latentpaint: Image inpainting in latent space with diffusion models. In *Win. App. Comput. Vis.*, pages 4334–4343, 2024. 1

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 33:6840–6851, 2020. 2, 3

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2, 3

[5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 2, 3

[7] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 3

[8] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, pages 9579–9589, 2024. 1, 2

[9] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *CVPR*, pages 10758–10768, 2022. 1

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 36:34892–34916, 2023. 1, 2, 3

[11] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *CVPR*, pages 21740–21751, 2024. 1, 3

[12] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, pages 11461–11471, 2022. 1

[13] Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. *arXiv preprint arXiv:2306.06101*, 2023. 3

[14] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 1, 2

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 3

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 1, 3

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

[18] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[19] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *CVPR*, pages 22428–22437, 2023. 1

[20] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. 2