2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Improving Robotic Manipulation Robustness via NICE Scene Surgery

Sajjad Pakdamansavoji* Mozhgan Pourkeshavarz* Adam Sigal* Zhiyuan Li
Rui Heng Yang Amir Rasouli

***Abstract*— Learning robust visuomotor policies for robotic manipulation remains a challenge in real-world settings, where visual distractors can significantly degrade performance and safety. In this work, we propose an effective and scalable framework, Naturalistic Inpainting for Context Enhancement (NICE). Our method minimizes out-of-distribution (OOD) gap in imitation learning by increasing visual diversity through construction of new experiences using existing demonstrations. By utilizing image generative frameworks and large language models, NICE performs three editing operations, object replacement, restyling, and removal of distracting (non-target) objects. These changes preserve spatial relationships without obstructing target objects and maintain action-label consistency. Unlike previous approaches, NICE requires no additional robot data collection, simulator access, or custom model training, making it readily applicable to existing robotic datasets.**

Using real-world scenes, we showcase the capability of our framework in producing photo-realistic scene enhancement. For downstream tasks, we use NICE data to finetune a vision-language model (VLM) for spatial affordance prediction and a vision-language-action (VLA) policy for object manipulation. Our evaluations show that NICE successfully minimizes OOD gaps, resulting in over 20% improvement in accuracy for affordance prediction in highly cluttered scenes. For manipulation tasks, success rate increases on average by 11% when testing in environments populated with distractors in different quantities. Furthermore, we show that our method improves visual robustness, lowering target confusion by 6%, and enhances safety by reducing collision rate by 7%.

I. Introduction

Robustness across visually diverse environments is fundamental for deploying robotic manipulation policies in the real world. Yet, learned policies, especially those trained via behavior cloning on demonstration datasets often suffer from significant performance and safety degradation when presented with visual distractors and scene variations to which they had no exposure during training [1]. To resolve the out-of-domain (OOD) learning gap, the trivial solution is to collect more data to cover the absent experiences, which can be very time-consuming and resource-intensive. To remedy this problem, some works have explored model-level solutions, such as object-centric representations [2], [3] and attention-guided policies [4], [5]. Others have made use of large-scale simulations, including physics based renderers, domain randomization techniques, and procedurally synthesized scenes, to diversify the training data [6]–[8]. However, such solutions are typically dependent on computationally expensive simulators and assume access to large-scale synthetic assets and rendering infrastructure, which are not readily available to practitioners.

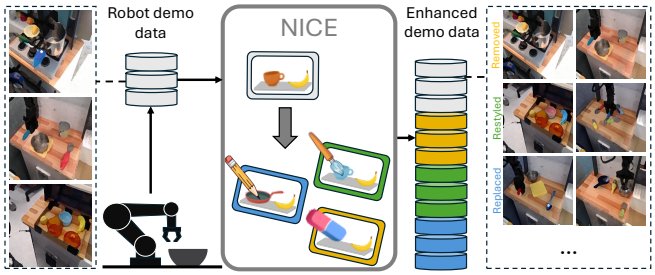

Fig. 1: An overview of our NICE generative framework. NICE uses the existing robot demonstration data and performs replacement, restyling, or removal operations on distracting objects to generate new experiences.

To this end, we propose Naturalistic Inpainting for Context Enhancement (NICE). NICE reduces the OOD gap by diversifying demonstration scenes to include varied visual distractors. More specifically, our NICE method applies in-place edits, including replacement with novel objects, restyling with new textures, and removal of distractors along with their shadows. The enhancements are done directly within real demonstration scenes, keeping the task, viewpoint, and target object fixed. To achieve this, NICE operates in two stages: scene decomposition in which objects and their corresponding masks are identified using an off-the-shelf vision-language-model (VLM); Role assignment and scene editing, in which distractor objects are identified, given the task description, and then modified using a combination of a diffusion generative model and language model, depending on the nature of the operation.

The enhanced data captures scene-level variability common in real-world environments that robots rarely encounter during training due to limited demonstration data. The NICE framework is compatible with any dataset of visual demonstrations and does not require modifications to the underlying robot hardware, control policies, or simulator infrastructure. We conduct evaluations to highlight the realism of our proposed framework, followed by two downstream tasks—visual spatial affordance prediction and object manipulation. We show that the NICE data can mitigate the negative effects of visual distractors on these tasks. In summary, the contributions of our work is as follow:

- We propose NICE, a novel framework for enhancing robotic data with scene-level surgery to improve the robustness of policies to distractors. Our framework effectively scales training data by increasing the vari-

*Equal Contribution. Corresponding author sajjad.pakdamansavoji@h-partners.com. Huawei Technologies Canada. *Work was done while at Huawei Canada.

ability of contextual elements with minimal human involvement.
- We evaluate the realism of NICE data against real-world examples in terms of both background consistency and overall quality of generation.
- We highlight the benefit of the NICE data on improving visual affordance prediction in scenes with various levels of clutter caused by distractors.
- Via extensive real-world examination, we validate how the NICE data can improve the robustness and safety of robotic manipulation policy across different tasks in environments populated with different numbers of distractors.

II. Related Works

Distractors in visual scene understanding. In the vision literature, distractors refer to visual elements that are irrelevant to the task at hand yet increase its complexity by diverting attention or introducing ambiguity [9]. These distractors may share visual properties with the target object (e.g., color, shape, or saliency), or they may differ in appearance but still act as confounding stimuli.

The impact of distractors has been widely studied across domains. In psychology, numerous studies have investigated how different types of distractors affect visual search [10], [11], as well as the role of attention mechanisms in mitigating their effects [12], [13]. Computer vision techniques have also been developed to address distractor-induced challenges, including category-level confusion in object detection [14], [15], and difficulties in distinguishing targets from visually similar distractors or handling occlusions in tracking tasks [16], [17].

In robotics, distractors similarly affect performance. For instance, in autonomous driving, a recent work based on the CausalAgents benchmark [18] shows that modifying irrelevant (non-causal) agents can substantially degrade prediction accuracy, prompting the need for causal reasoning approaches [19], [20]. For robot localization, the authors of [21] demonstrate that salient distractors can disrupt performance and propose a technique to suppress their influence. Distractors in cluttered environments are shown to impact robotic manipulation by interfering with object recognition and complicating grasping actions [22]–[28]. In some cases, these distractors not only obscure the target but also lead to incorrect action generation. For example, the study in [29] shows that simply altering distractors, either by replacing them with similar items of different color or by swapping them entirely, can reduce the policy's success rate by as much as 50% across multiple manipulation benchmarks.

Visual Robustness in Robotic Manipulation. Visuomotor policies for robotic manipulation, particularly those trained via behavior cloning, are known to be sensitive to distribution shifts in the visual input space. Prior work has addressed this limitation through architectural enhancements, such as object-centric representations [30], and multimodal foundation models like RT-1 [31], which leverage large-scale data and temporal context to improve robustness. Other approaches such as ImitDiff [32] incorporate semantic segmentation at inference time to guide the policy toward task-relevant regions, demonstrating improved performance under cluttered or distractor-heavy conditions. While these methods improve robustness, they often require complex perception modules or large-scale training infrastructure. Our work instead focuses on improving robustness through data-centric augmentation applied directly to real-world visual demonstrations.

Data Augmentation in Robotics. Domain randomization [33], [34] has long been used to train visual policies in simulation by exposing models to randomized textures, lighting, and object appearances. However, these techniques are limited to sim-to-real transfer and are rarely applied to real demonstration data. More recently, augmentation methods that directly operate on real robot datasets have shown promise. RoboSaGA [35] employs saliency-guided background replacement using out-of-domain images to preserve task-relevant content while introducing variability. ROSIE [36] uses diffusion models to semantically edit scenes by adding or replacing objects, enhancing generalization to unseen configurations. The method in [37] combines generative image editing with 3D object rendering to generate hundreds of semantically diverse distractor variants per scene. These methods typically rely on large-scale or proprietary generative models or on simulation assets, which may introduce domain gaps or require significant compute, whereas our approach uses an open, reproducible generative component and avoids proprietary dependencies. In addition, they lack studies to showcase their level of realism which can be a source of sim2real gap. In contrast, our method uses direct visual editing to inject distractors into real images with minimal overhead, enabling simple augmentation for existing datasets. We show that our method is effective for generating realistic scenes, resulting in better real-world adaptation. Furthermore, our method performs backward operation by removing distractors. This operation can better enforce the policy to learn the effect of non-targets, and the difference between cluttered and non-cluttered scenes.

Synthetic Scene Editing. Beyond robotics, recent works have explored scene editing as a means to evaluate and improve model robustness across a range of domains, underscoring the growing importance of visual generalization. For action recognition, HAT [38] uses video inpainting to isolate or remove humans from real-world clips, revealing strong biases toward background features. In the context of image authenticity detection, Semi-Truths [39] introduces localized AI-generated edits into real photographs to test detector sensitivity to subtle manipulations. UltraEdit [40] leverages diffusion-based region editing to build a large-scale benchmark for instruction-based image manipulation. Collectively, these efforts highlight scene alteration as a powerful, domain-agnostic strategy for analyzing and enhancing model robustness. Our work brings this perspective to robotic manipulation, where sensitivity to visual distractors remains a key challenge. By adapting real-world scene editing techniques to robot demonstration data, we contribute

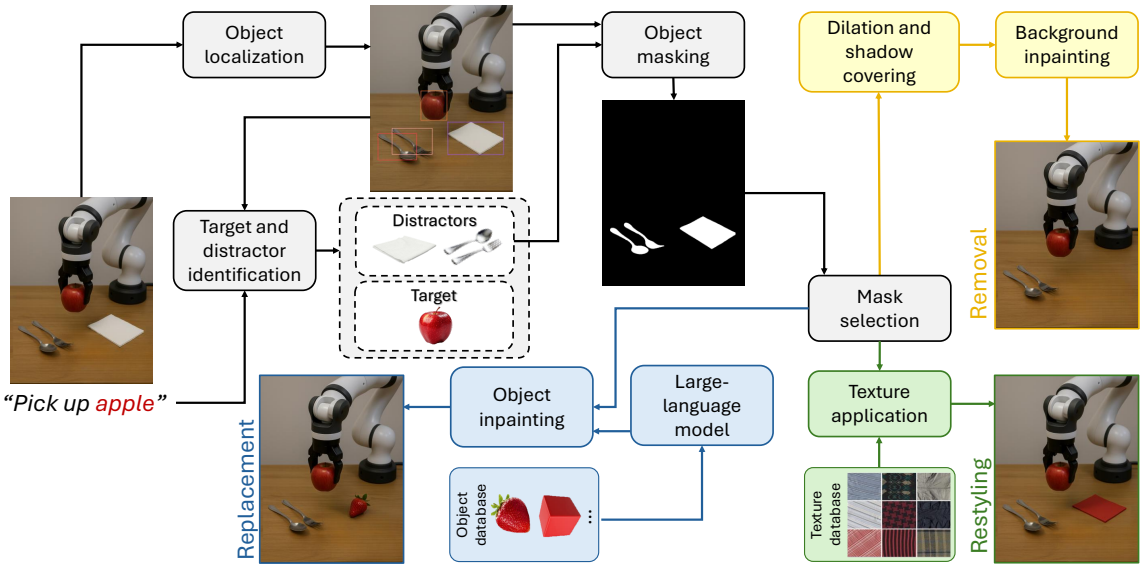

Fig. 2: An overview of the NICE framework. The method starts by detecting all objects, identifying the target, and segmenting distracting objects. The object of interest is then selected to perform one of the following operations. **Removal** dilates the selected mask to cover shadows and feeds it into a inpainting model to fill with background texture. **Restyling** uses a texture database and applies it to the selected mask to change the appearance of the distractor. **Replacement** uses a large-language model to generate object description, which is then fed into an image inpainting module to replace the distractor.

both a practical data enhancement pipeline and an evaluation framework grounded in physical robot experiments.

III. Methodology

A. Problem Setup

We consider a standard behavioral cloning setup for visuomotor object manipulation tasks, in which a robot observes RGB images of the scene and outputs corresponding manipulation actions. The training data consist of demonstrations, where each sample includes an image observation, the robot arm's state, the action executed by the expert policy, and an associated task instruction. The objective is to learn a visuomotor policy that, conditioned on the task instruction, maps observations to actions that imitate the demonstrated behavior. Our aim is to enhance the robustness of policy learning in the presence of visual distractors by enriching the training data with diverse and systematically varied distractor instances while preserving the original task semantics.

B. Overview of NICE

NICE takes real demonstration images and applies diverse scene enhancements to simulate novel visual distraction, thus generating additional training data. Our framework performs three types of edits: **removal**, **replacement**, and **restyling** of distractors while keeping the original target object and its relation to the demonstration unchanged. A key design principle is action-label consistency, meaning that after enhancement, the image should still correspond to the same grasp or manipulation action as before. To this end, we do not delete or occlude the target object. As for non-target objects, we only change the configuration by replacement or removal not adding new objects. Although adding new object can further enhance the scene by, for example, creating occlusions, it can potentially impact the planned trajectory or perhaps affect the realism as scaling the object correctly can be challenging. We further insure that the new instances of the inserted distractors do not conflict with the recorded trajectory. For this, we enforce the new object to be within similar dimensions as the original one via instruction to the generative model. In other words, the task-relevant causal features (e.g. the block to pick up) are invariant under the edits. To preserve the realism of the generated images, we create separate versions of the same image using one of the three editing operations. As shown in Figure 2, the pipeline consists of two stages: Scene decomposition and role assignment, and scene editing.

C. Scene Decomposition and Role Assignment

Object Parsing. The first step is to detect all objects in the scene. We use Florence-2 [41], a multitask VLM that detects objects with or without text prompts. Florence-2 produces bounding boxes and class labels for each object. The bounding boxes are then passed to the Segment Anything model v2 (SAM-2) [42] to compute precise segmentation masks, along with confidence scores for each object (see Figure 3 for an example).

Target and Distractor Identification. It is important to accurately distinguish between the target and distractors. Given a task instruction (e.g. pick up the blue cube), we identify the target among all detected objects. Using the predicted classes generated by object parsing step, we exclude the target from the segmentation operation. In addition, to improve the consistency of the scenes (e.g. avoid major artifacts in the scene), we exclude very large objects, whose bounding boxes' dimensions exceed 40% (set empirically)

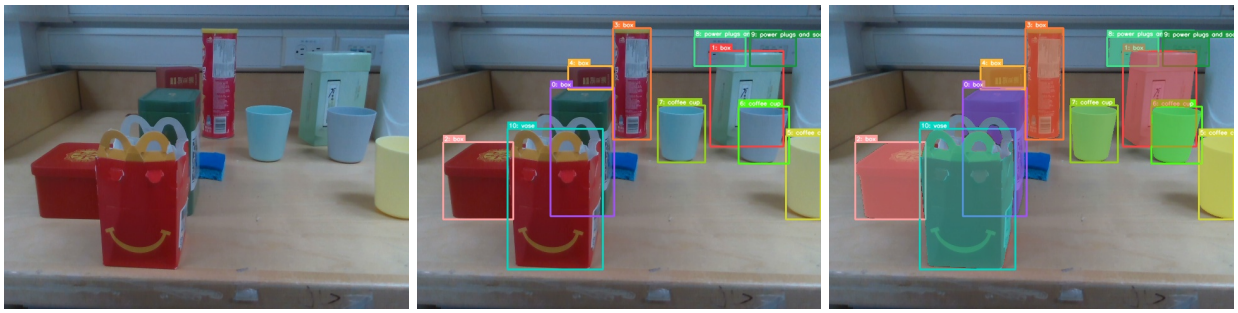

Fig. 3: Examples of the object parsing step: (**Left**) input raw image, (**Middle**) object detection results using Florence-2, and (**Right**) segmentation results using SAM-2.

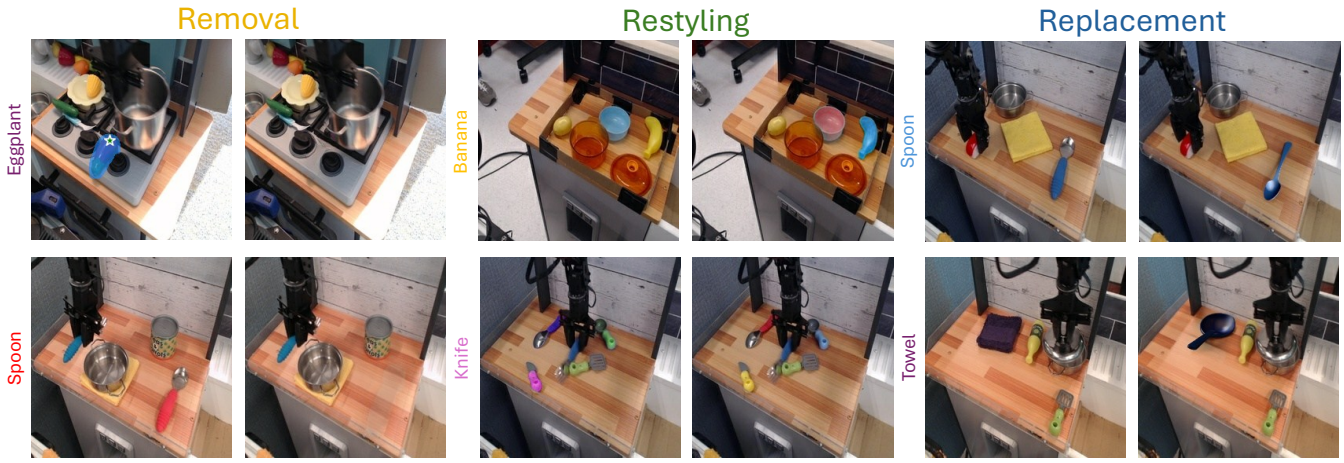

Fig. 4: Examples of data enhancement using NICE on the Bridge data [43]. In each image pair, left is the original image and right is the edited one. Name and color of the manipulated object is mentioned on the left side of each image pair.

of the image height or width. All other remaining objects are considered as potential candidates for editing.

D. Scene Editing

For each candidate distractor object, NICE performs one of three edit operations on the copies of the original images (see Figure 4 for an example). The operations are performed as follows:

Object Removal. For a given image, a random set of 0 to n object masks are chosen and combined into a single mask for removal (where n is the number of objects, excluding large size ones or the target). This mask is then dilated with a hyperparameter dil (set empirically) to smooth the edges and cover the original object's shadow. Finally, we mask out the combined distractor region and apply the LaMa inpainting model [44] to fill it with background content. LaMa is a large-mask image inpainting model based on Fourier convolutions. It propagates texture from surrounding pixels to plausibly reconstruct the scene.

Object Restyling. Our goal is to change the appearance, texture, or color of an object without altering its shape or pose. For this, we follow the same masking strategy as in removal, generating n masks. Then, we sample textures from the Describable Textures Dataset (DTD) [45], which contains thousands of real texture patches (e.g. dotted, striped, etc.) applicable to object surfaces. We project the texture onto the object mask by overlaying and adjusting color or by performing stylization. For example, a wooden block might be recolored with a zebra pattern or a metallic spoon with a rust texture. The color and appearance of the objects are altered by adjusting their brightness, hue, and saturation. These transformations are applied to the object masks to introduce controlled variability in visual attributes.

Object Replacement. Unlike object removal and restyling, for each replacement operation, we exchange one object at a time. To maintain realism and consistency, we replace each object with another object that is congruent within the given scene. More specifically, after masking out the target region of the image along with dilation, we use the Stable Diffusion inpainting model [46] to generate the recommended object via a structured prompt containing the name of the new object. For example, caption might say "a yellow block on a wooden table", and the diffusion model synthesizes the block with appropriate lighting. This insertion leverages state-of-the-art generative priors to produce photo-realistic novel objects.

For replacement, we can employ two different strategies. 1- Generate an instance of the same object category with different appearance, by passing its name to the diffusion model and ask to alter it. 2- Generate a context relevant yet visually distinct object from other categories (e.g. replacing

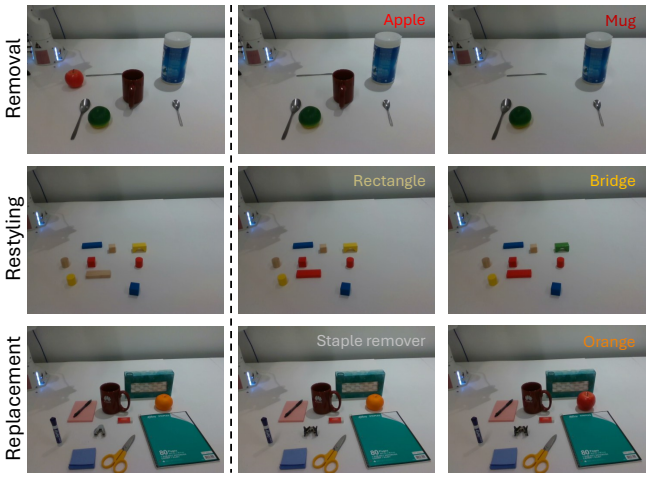

Fig. 5: Real-world replication of editing operations used for validation of the realism of the NICE data. For each series of samples, the scene was populated with multiple objects. Then one at the time, each object (indicated on the top-right corner of each image) was either removed, replaced with the same object with different color, or replaced with another object entirely.

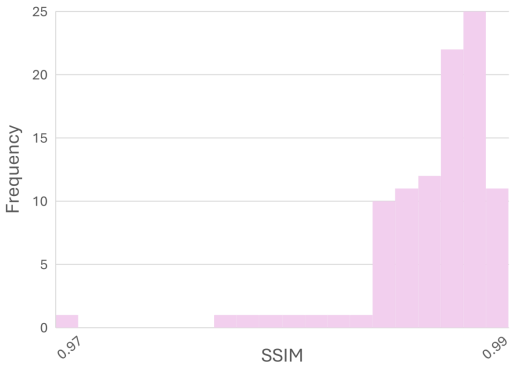

Fig. 6: Distribution of SSIM values for removal operation on real-world data using NICE.

a graphite cooking pan with a gray dish cloth as shown in Figure 4). This allows us to generate a novel scene while maintaining the context. For this purpose, we use Deepseek-r1:7b [47] via the Ollama framework [48] to generate a description of a household object similar in size to the original one, which is then fed into the Stable Diffusion model [46]. In our experiments, we found using such a small language model suffices for accurate prompting in order to generate similar objects.

IV. Evaluation

A. Background Consistency

One of the key considerations for scene editing is to maintain background consistency. This is especially challenging when removing an object, since the background must be reconstructed and secondary effects, such as shadows, must also be eliminated. Here, we examine the ability of our method to achieve this goal in the case of removal. For this, we create 20 cluttered scenes in real world. We then capture 5 variations of the scenes by removing one object at a time, for a total of 100 real-world images (see example in Figure 5). We then replicate these changes using our pipeline and compare to real images using the Structural Similarity Index Measure (SSIM) metric [49]. This metric computes the similarity of the generated image based on a weighted average of three components, namely luminance, contrast, and structure. As shown in Figure 6, our method generally yields a very high score on generated samples, indicating its accuracy in reconstructing the background.

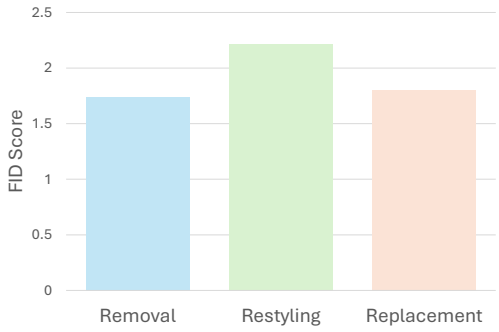

Fig. 7: FID score of the three enhancement strategies on real-world images.

B. Data Generation Realism

Following the similar procedure as in IV-A, we capture real-world images for restyling and replacement. For the former operation, we swap the objects with the same objects of different color and for latter, with objects of similar category (e.g. orange with an apple). The real-world data samples are shown in Figure 5. Using our framework, we then replicate the scene alterations and compute Fréchet Inception Distance (FID) [50] between the generated and real-world captured images. FID is computed by measuring the Fréchet Distance between probability distributions of reference and generated images based on Inception network embeddings [51]. As shown in Figure 7, lower FID scores indicate that our enhanced images perceptually and statistically are close to the real images. The higher FID value of restyling can be due to the fact that generative models are more successful at modeling ambient conditions (e.g. lighting) when generating an entire object as opposed to restyling the texture of an existing object.

C. Spatial Affordance for Robotics Manipulation

One of the key issues caused by distractors is visual confusion, which diminishes the ability of the robot to accurately localize the target object and identify affordance regions for performing manipulation. We employ RoboPoint [52], a state-of-the-art vision-language-model (VLM) that predicts spatial affordance in free space, which then can be used for any downstream robotic task.

For this experiment, as shown in Figure 8, we consider scenes with three levels of clutter: *low clutter (LC)* with 1-2 objects, *medium clutter (MC)* with 5-8 objects, and *high*

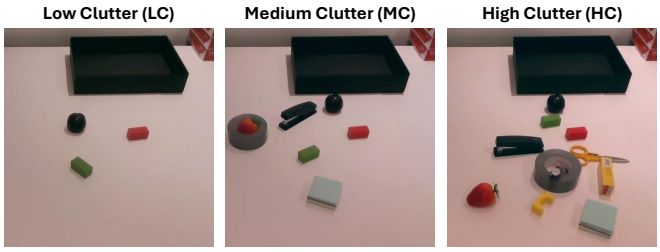

Fig. 8: Samples of scenes with different levels of clutter.

TABLE I: Average prediction accuracy (APA)(%) across different clutter levels using RoboPoint [52].

Dataset	APA_{LC}	APA_{MC}	APA_{HC}
Original	32.64	30.47	20.08
+NICE	**48.12** (+15.48)	**45.76** (+15.29)	**41.44** (+21.36)

clutter (HC) containing 11-15 objects. In every scene, we insert at least one distractor that is visually or semantically similar to the target, as well as additional distractors that differ in category, geometry, or appearance. In high clutter, the objects are densely placed to increase difficulty. Following the protocol in [52], we report the results using *average prediction accuracy (APA)*, which measures the percentage of predicted points that fall within the ground-truth target mask.

As the results in Table I suggest, our enhancement method can significantly improve the affordance prediction performance. In low and medium clutter scenes we observe an increase of more than 15% in APA, reaching up to 21% in high cluttered scenes. This emphasizes the challenge distractors can pose to robot's perception as clutter level increases. Scaling the data using NICE can greatly compensate for such degradation and result in more stable performance across scenes with different levels of clutter.

D. Robotic Manipulation in Clutter

In this experiment, we directly measure the impact of the data generated using NICE on the downstream object manipulation tasks. We adopt the four core skills from [53], namely **pick** object, **move** one object close to another, **put** one object on another, and **stack** two objects. We design our experimental setup with 6 levels of clutter using 0,1,2,4,8, and 16 distracting objects in the scene. For each setup, we consider 9 different variants (a grid of 3 x 3) for each skill by changing the position of the target object(s). In total, we evaluate on 216 scenarios. Sample scene configurations are shown in Figure 9.

As our manipulation policy, we choose π_0 [54], a vision-language-action model pretrained on Open X-Embodiment [55]. We finetune four versions of the policy using the following four datasets; **Base** data contains 42 demonstrations for each skill collected using only the target objects without any distractors; **8-Dist** contains data with only 8 distractors with 9 variations for each of the four skills; **Full** comprises 45 real demonstrations for each level of clutter for each skill. **NICE** consists of enhanced data generated using the 8-Dist data as input. For each skill, we create 54 samples by applying 2 variations of each of the 3 enhancement operations. The first model is only finetuned on Base data and subsequent ones are finetuned on Base plus one of the three distracted datasets. We evaluate each finetuned model on all test scenarios, a total of 864 real-world trials.

NICE data improves performance at different levels of clutter. We begin by measuring the success rate of the policy at different levels of clutter. The results, averaged over all skills, are shown in Figure 9. Here, we can see that compared to Base and 8-Dist, NICE data significantly boost the performance of the policy, especially on scenes populated with more distractors. Even though the enhancement is performed on scenes with 8 distractors, the benefit extends to more complex scenes as well, thanks to the diversification of the distractors. Overall, the NICE policy's SR improves upon 8-Dist (which used as input for enhancement) by 11% and is on par with Full, showing how our method can achieve the same performance without any exposure to manually crafted scenarios.

NICE data lowers different types of failures. We are interested to determine whether NICE can lower failures, especially those due to visual scene clutter. For this we consider collision rate (CR) measuring the percentage of scenarios in which the arm makes contact with a distractor and target confusion rate (TCR) capturing the percentage of scenarios in which the robot reaches for a non-target object for grasping. As shown in Table II, NICE significantly lowers collision rate and target confusion by 7% and 6% compared to 8-Dist (which used as seed data for enhancement) and by 12% and 4% compared to Full. Using the same distractors in different settings in manually crafted data, as in 8-Dist, has a positive effect, lowering CR potentially by 16%. However, additional samples in the Full dataset results in more overfitting, reducing the advantageous gap. NICE, on the other hand, diversifies the data, leading to improvement of the collision rate even without changing the location of the distractors in the scene. The effect of diversification is also apparent in TCR, allowing the policy to better learn the distinctive features of the target objects.

TABLE II: The average performance of the policy finetuned with each dataset. Abbreviations stand for success rate (SR), collision rate (CR), and target confusion rate (TCR). The direction of arrows indicate higher or lower values are better.

	SR↑	CR↓	TCR↓
Base	47	38	34
+8-Dist	53	22	16
+Full	**65**	27	14
+NICE	64	**15**	**10**

Despite lowering CR and TCR significantly, NICE achieves similar SR to Full. There are two main reasons for

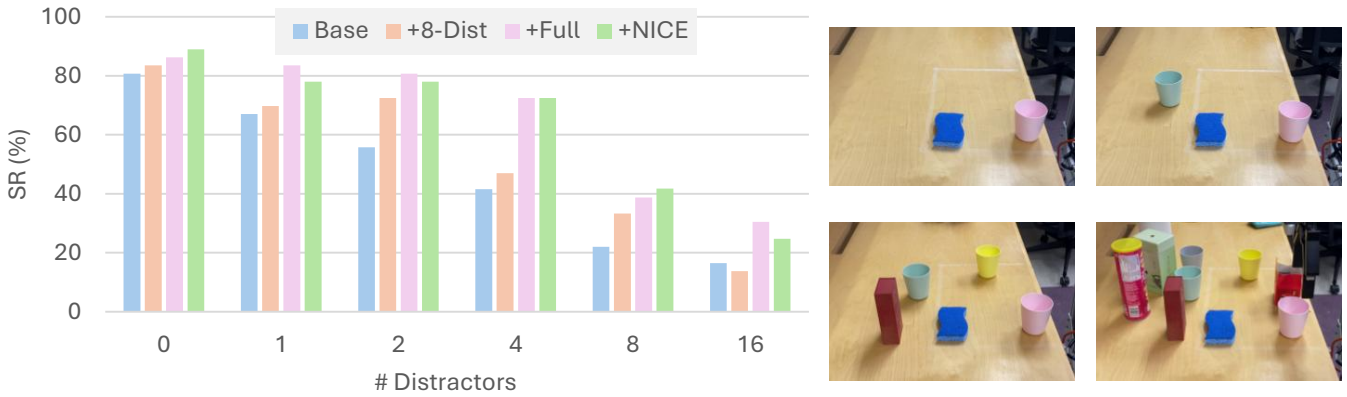

Fig. 9: (**Left**) Performance of the manipulation policy π_0, finetuned on three data configurations. The results are averaged over all skills (**Right**) Example experiment scenes with varying numbers of distractors.

TABLE III: The performance of the policy average per skill. Results are shown as success rate (SR)/collision rate (CR). For the former higher is better and latter lower.

	Pick	**Move**	**Put**	**Stack**	**Average**
Base	48/46	52/30	44/39	45/37	47/38
+8-Dist	59/13	56/28	50/37	48/11	53/22
+Full	**80**/20	65/19	58/54	**59**/17	**65**/27
+NICE	67/**9**	**67/20**	**63/26**	**59/5**	64/**15**

this: 1- as per our evaluation, we do not consider collision as a failure as long as the task is completed. Hence, NICE achieves similar level of success but more safely. 2- There are other types of failures, e.g. grasp failure, placement failure, etc. which can contribute to lack of success. For instance, since in NICE we do not alter the position of the target objects, their poses are not diversified, hence no added advantage is achieved. The Full data, on the other hand, contains scenes with different setups for the targets, leading to more robust grasping.

NICE benefit amplifies as the tasks get more complex We consider the per-task breakdown of the results in Table III. Here, we can see that the more complex the tasks get, the more benefit is gained using the NICE data. In the Pick task, which only involves lifting the target, the NICE policy lags behind in SR due to lower advantage in improving grasp robustness (as discussed in the previous experiment). However, it still drastically lowers collision rate. The reduction in CR extends to more complex tasks, where the gap with Full becomes even higher, 28% in Put and 12% in Stack tasks. The reason for this is that in these tasks two targets are involved, the object that is to be grasps and the one on which the target is to be placed on. Hence manipulating one with respect to the other, increases the likelihood of collision with distractors. Overall, compared to the Base data and the 8-Dist data, NICE clearly stands out on all metrics for all tasks.

V. Conclusion and Future Work

In this work, we proposed a novel approach for enhancing robot data without the need for action generation or human involvement. Our NICE method, relies on a language conditioned generative model to identify the objects of interest and performs scene editing by either removing, restyling, or replacing distractors with novel objects or background. Through empirical evaluation on real-world data, we showed that our pipeline generates realistic scenes that significantly improve robot perception and, consequently, downstream manipulation tasks.

In this work we mainly focused on three forms of scene enhancement, namely removal, restyling, and replacement. We argued that correct selection of novel objects can maintain the realism of the scenarios, e.g. not obscuring robot movement in the pre-recorded scenes. For data generation, other forms of enhancement can be considered, such as rearrangement or addition. However, these operations require better understanding of robot actions in the 3D space to maintain realism. We will consider such extensions for our future work.

In this work, we examined the impact of our framework on spatial affordance prediction and core manipulation tasks. It is reasonable to assume that visual confusion or operational confinement caused by distractors can have different degrees of impact on different manipulation tasks. For example, relative to object-picking tasks, object arrangement poses greater challenges, as it increases the likelihood of confusion. We plan to extend our empirical evaluation on a large set of robotic skills to both identify challenges posed by distractors and clutter and determine whether our proposed data enhancement framework can be used to mitigate them.

References

[1] W. Pumacay, I. Singh, J. Duan, R. Krishna, J. Thomason, and D. Fox, "The colosseum: A benchmark for evaluating generalization for robotic manipulation," in *RSS*, 2024.

[2] A. Chapin, B. Machado, E. Dellandrea, and L. Chen, "Object-centric representations improve policy generalization in robot manipulation," *arXiv preprint arXiv:2505.11563*, 2025.

[3] W. Yuan, C. Paxton, K. Desingh, and D. Fox, "Sornet: Spatial object-centric representations for sequential manipulation," in *CoRL*, 2022.

[4] J. Zhang, Y. Gu, J. Gao, H. Lin, Q. Sun, X. Sun, X. Xue, and Y. Fu, "Lac-net: Linear-fusion attention-guided convolutional network for accurate robotic grasping under the occlusion," in *IROS*, 2024.

[5] S. James, K. Wada, T. Laidlow, and A. J. Davison, "Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation," in *CVPR*, 2022.

[6] Y. Wang, Z. Xian, F. Chen, *et al.*, "Robogen: towards unleashing infinite data for automated robot learning via generative simulation," in *ICML*, 2024.

[7] Z. Zhou, P. Atreya, A. Lee, H. R. Walke, O. Mees, and S. Levine, "Autonomous improvement of instruction following skills via foundation models," in *CoRL*, 2024.

[8] C. R. Garrett, A. Mandlekar, B. Wen, and D. Fox, "Skillmimicgen: Automated demonstration generation for efficient skill learning and deployment," in *CoRL*, 2024.

[9] H. R. Liesefeld, D. Lamy, N. Gaspelin, J. J. Geng, D. Kerzel, J. D. Schall, H. A. Allen, B. A. Anderson, S. Boettcher, N. A. Busch, *et al.*, "Terms of debate: Consensus definitions to guide the scientific discourse on visual distraction," *Attention, Perception, & Psychophysics*, vol. 86, no. 5, pp. 1445–1472, 2024.

[10] B. Olk, A. Dinu, D. J. Zielinski, and R. Kopper, "Measuring visual search and distraction in immersive virtual reality," *Royal Society Open Science*, vol. 5, no. 5, p. 172331, 2018.

[11] M. A. Petilli, F. Marini, and R. Daini, "Distractor context manipulation in visual search: How expectations modulate proactive control," *Cognition*, vol. 196, p. 104129, 2020.

[12] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, no. 1-2, pp. 507–545, 1995.

[13] L. Chelazzi, F. Marini, D. Pascucci, and M. Turatto, "Getting rid of visual distractors: The why, when, how, and where," *Current Opinion in Psychology*, vol. 29, pp. 135–147, 2019.

[14] Y. Li, H. Zhu, Y. Cheng, *et al.*, "Few-shot object detection via classification refinement and distractor retreatment," in *CVPR*, 2021.

[15] Y.-C. Liu, C.-Y. Ma, X. Dai, J. Tian, P. Vajda, Z. He, and Z. Kira, "Open-set semi-supervised object detection," in *ECCV*, 2022.

[16] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *ECCV*, 2018.

[17] F. Zhong, P. Sun, W. Luo, T. Yan, and Y. Wang, "Towards distraction-robust active visual tracking," in *ICML*, 2021.

[18] L. Sun, R. Roelofs, B. Caine, K. S. Refaat, B. Sapp, S. Ettinger, and W. Chai, "Causalagents: a robustness benchmark for motion forecasting," in *ICRA*, 2024.

[19] M. Pourkeshavarz, J. Zhang, and A. Rasouli, "Cadet: a causal disentanglement approach for robust trajectory prediction in autonomous driving," in *CVPR*, 2024.

[20] E. Ahmadi, R. Mercurius, S. Alizadeh, K. Rezaee, and A. Rasouli, "Curb your attention: Causal attention gating for robust trajectory prediction in autonomous driving," in *ICRA*, 2025.

[21] O. Mendez, M. Vowels, and R. Bowden, "Improving robot localisation by ignoring visual distraction," in *IROS*, 2021.

[22] N. Di Palo and E. Johns, "Keypoint action tokens enable in-context imitation learning in robotics," in *RSS*, 2024.

[23] H. Kim, Y. Ohmura, and Y. Kuniyoshi, "Using human gaze to improve robustness against irrelevant objects in robot manipulation tasks," *RAL*, vol. 5, no. 3, pp. 4415–4422, 2020.

[24] H. Kim, Y. Ohmura, and Y. Kuniyoshi, "Transformer-based deep imitation learning for dual-arm robot manipulation," in *IROS*, 2021.

[25] E. U. Samani and A. G. Banerjee, "Persistent homology meets object unity: Object recognition in clutter," *Transactions on Robotics*, vol. 40, pp. 886–902, 2024.

[26] H. Kasaei, M. Kasaei, G. Tziafas, S. Luo, and R. Sasso, "Simultaneous multi-view object recognition and grasping in open-ended domains," *Journal of Intelligent & Robotic Systems*, vol. 110, no. 2, p. 62, 2024.

[27] B. Tang and G. S. Sukhatme, "Selective object rearrangement in clutter," in *CoRL*, 2023.

[28] A. Ummadisingu, K. Takahashi, and N. Fukaya, "Cluttered food grasping with adaptive fingers and synthetic-data trained object detection," in *ICRA*, 2022.

[29] S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh, and P. Liang, "Language-driven representation learning for robotics," in *RSS*, 2023.

[30] Y. Zhu, Z. Jiang, P. Stone, and Y. Zhu, "Learning generalizable manipulation policies with object-centric 3d representations," in *CoRL*, 2023.

[31] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," in *RSS*, 2023.

[32] Y. Dong, H. Ge, Y. Zeng, J. Zhang, B. Tian, G. Tian, H. Zhu, Y. Jia, R. Wang, R. Yi, *et al.*, "Imit diff: Semantics guided diffusion transformer with dual resolution fusion for imitation learning," *arXiv preprint arXiv:2502.09649*, 2025.

[33] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IROS*, 2017.

[34] F. Sadeghi and S. Levine, "Cad2rl: Real single-image flight without a single real image," in *RSS*, 2017.

[35] Z. Zhuang, R. Wang, N. Ingelhag, V. Kyrki, and D. Kragic, "Enhancing visual domain robustness in behaviour cloning via saliency-guided augmentation," in *CoRL*, 2024.

[36] T. Yu, T. Xiao, A. Stone, J. Tompson, A. Brohan, S. Wang, J. Singh, C. Tan, J. Peralta, B. Ichter, *et al.*, "Scaling robot learning with semantically imagined experience," *arXiv preprint arXiv:2302.11550*, 2023.

[37] Z. Chen, Z. Mandi, H. Bharadhwaj, M. Sharma, S. Song, A. Gupta, and V. Kumar, "Semantically controllable augmentations for generalizable robot learning," *IJRR*, 2024.

[38] J. Chung, Y. Wu, and O. Russakovsky, "Enabling detailed action recognition evaluation through video dataset augmentation," in *NeurIPS*, 2022.

[39] A. Pal, J. Kruk, M. Phute, M. Bhattaram, D. Yang, D. H. Chau, and J. Hoffman, "Semi-truths: A large-scale dataset of ai-augmented images for evaluating robustness of ai-generated image detectors," in *NeurIPS*, 2024.

[40] H. Zhao, X. S. Ma, L. Chen, S. Si, R. Wu, K. An, P. Yu, M. Zhang, Q. Li, and B. Chang, "Ultraedit: Instruction-based fine-grained image editing at scale," in *NeurIPS*, 2024.

[41] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan, "Florence-2: Advancing a unified representation for a variety of vision tasks," in *CVPR*, 2024.

[42] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollar, and C. Feichtenhofer, "SAM 2: Segment anything in images and videos," in *ICLR*, 2025.

[43] H. Walke, K. Black, *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *CoRL*, 2023.

[44] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *WACV*, 2022.

[45] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi, "Describing textures in the wild," in *CVPR*, 2014.

[46] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[47] D. Guo, D. Yang, H. Zhang, *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[48] F. S. Marcondes, A. Gala, R. Magalhães, F. Perez de Britto, D. Durães, and P. Novais, "Using ollama," in *Natural Language Analytics with Generative Large-Language Models: A Practical Approach with Ollama and Open-Source LLMs*, 2025, pp. 23–35.

[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[50] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NeurIPS*, 2017.

[51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

[52] W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, and D. Fox, "Robopoint: A vision-language model for spatial affordance prediction in robotics," in *CoRL*, 2024.

[53] X. Li, K. Hsu, J. Gu, O. Mees, K. Pertsch, and other, "Evaluating real-world robot manipulation policies in simulation," in *CoRL*, 2024.

[54] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "pi0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[55] A. O'Neill, A. Rehman, A. Maddukuri, Gupta, *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0," in *ICRA*, 2024.