Don't Get Distracted: Improving Robotic Perception Robustness via In-Context Visual Scene Editing

Mozhgan Pourkeshavarz*, Adam Sigal*, Sajjad Pakdamansavoji,
Zhiyuan Li, Rui Heng Yang, Amir Rasouli**
*Equal Contriution
Huawei Technologies Canada
**amir.rasouli@huawei.com

Abstract: Learning robust visuomotor policies for robotic manipulation remains a challenge in real-world settings, where visual distractors and clutter can significantly degrade performance. In this work, we highlight the challenges that visual clutter poses to robotic manipulation and propose an effective and scalable i**n**-context v**i**sual s**c**ene **e**diting (**NICE**) strategy based on real-world images. Our method synthesizes new variations of existing robot demonstration datasets by programmatically modifying non-target objects directly within the real scenes. This approach diversifies environmental conditions without requiring additional action generation, synthetic rendering, or simulator access. Using real-world scenes, we showcase the capability of our framework in performing realistic object replacement, restyling, and removal. We generate new data using NICE and finetune a vision-language model (VLM) for spatial affordance and a vision-language-action (VLA) policy for object manipulation. Our experiments show that using our editing framework results in more than a 20% increase in both accuracy in affordance prediction and success rate in manipulation.

Keywords: Data enhancement, Visual clutter, Affordance prediction, Manipulation

1 Introduction

Generalization 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 degradation when presented with visual distractors, background clutter, or other scene variations not encountered during training [1]. Recent work has attempted to mitigate this problem by exploring model-level solutions, such as object-centric representations [2, 3] and attention-guided policies [4, 5]. In parallel, large-scale simulation pipelines have enabled domain randomization and synthetic data generation that diversify training data [6, 7, 8]. However, such solutions are either dependent on complex perception modules, computationally expensive simulators, or assume access to large-scale synthetic assets and rendering infrastructure. In contrast, relatively little attention has been given to simple, scalable, and data enhancement methods that operate directly on real-world visual scenes.

In this work, we propose a targeted i**n**-context v**i**sual s**c**ene **e**diting (**NICE**) strategy. Our method edits real demonstration scenes by modifying distractor elements, such as objects of varied color, shape, and texture, and background clutter, directly within real images. These edits simulate the type of visual variability robots commonly face in everyday real-world environments, but are rarely exposed to during training. Our 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 for data editing. This is followed by assessing the impact of our method on two downstream tasks—visual spatial affordance prediction and object manipulation. We show that our NICE strategy can mitigate the negative effects of visual distractors on these tasks.

2 Related Works

Distractors in visual scene understanding. In the vision literature, distractors refer to visual elements that are irrelevant to the task at hand. Distractors have been shown to increase the complexity of the task by diverting attention or introducing ambiguity [9]. Some distractors share the target's visual features (color, shape, texture). Others are visually unrelated but still cause mislocalization or false positives by diverting attention. The impact of distractors has been widely studied across different 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]. In computer vision, techniques have 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, recent work based on the CausalAgents benchmark [18] showed that modifying irrelevant (non-causal) agents can substantially degrade prediction accuracy, prompting the need for causal reasoning approaches [19, 20]. In robotic manipulation, distractors in cluttered environments can interfere with object recognition and grasp pose estimation [21, 22, 23, 24, 25, 26, 27]. In some cases, these distractors not only obscure the target but also lead to incorrect action generation by causing confusion [28, 1]. For example, [29] report that simply altering context, either by replacing the non-target objects with visually similar, color-variant lookalikes or entirely different objects, can reduce policy success rates by as much as 50%. The authors of [1] empirically showed how environmental factors impact manipulation policies. They found that distractors and contextual elements, such as lighting, camera pose, and target characteristics can significantly hinder performance. One way to mitigate the negative impact of distractors is to expose the policy to additional diverse data during training. To achieve this goal, We propose a novel method that effectively diversifies data in an automated and scalable manner.

Data augmentation in robotics. Domain randomization [30, 31] has long been used to train visual policies in simulation by exposing models to randomized textures, lighting, and object appearances. However, the effectiveness of simulated randomization is limited by sim-to-real transfer. Recent works remedy this issue by training and augmenting directly on real robot data. RoboSaGA [32] replaces the background using out-of-domain images to preserve task-relevant content while introducing variability. ROSIE [33] uses diffusion models to edit scenes by adding or replacing objects (often of similar category or shape), enhancing generalization to unseen configurations. In [34], the authors combine generative image editing with 3D object rendering to generate hundreds of diverse distractor variants per scene. However, using generative models requires significant compute and simulation assets introduce domain gaps due to lack of realism. In contrast, our method uses direct visual editing to modify distractors in real images with minimal overhead, achieving high degree of realism and enabling object modification across different objects categories with varied shape and form. Using an effective language-guided mechanism, our framework automatically and at scale enhances any existing datasets while eliminating the need for new demonstrations.

3 Methodology

3.1 Problem Setup

We consider a standard behavioral cloning setup for a visuomotor pick-and-place task, 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 replicate 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.

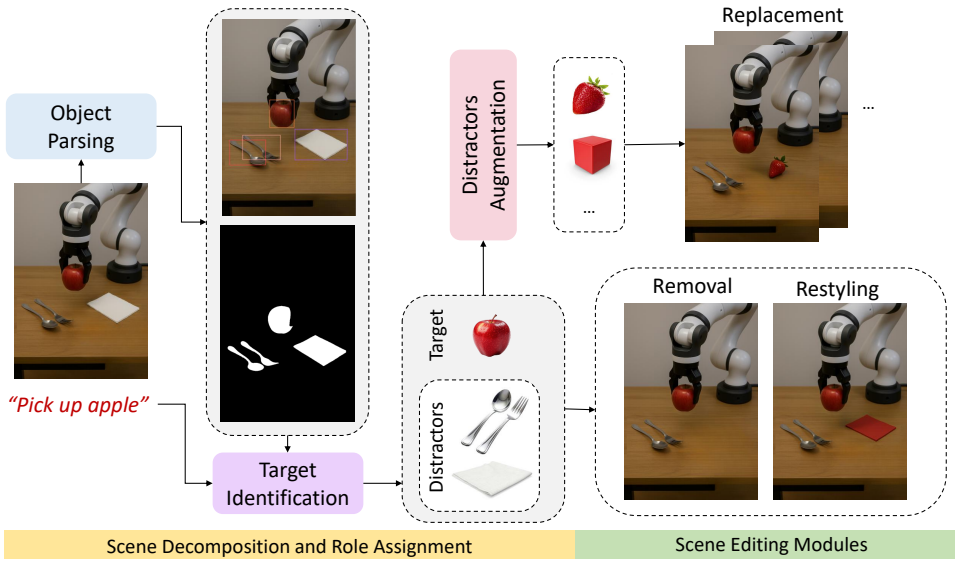

Figure 1: Overview of the proposed in-context visual scene editing (NICE) pipeline. It starts by parsing and masking the distractor (non-target) objects. Then depending on the enrichment strategy (removal, restyling, or replacement), the given object is exchanged with a new object or background.

3.2 Overview of NICE

NICE takes real demonstrations and applies diverse scene enhancements to simulate novel visual clutter, thus extending training data. NICE 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 pick-and-place action as before. To this end, we do not delete or occlude the target object. We further insure that the new instances of the inserted distractors do not conflict with the recorded trajectory. In other words, the task-relevant causal features (e.g. the block to pick up) are invariant under the augmentation. In practice, we randomly perform one of the three edits per image to produce a varied augmented dataset. As shown in Figure 1, the pipeline consists of two stages: Scene decomposition and role assignment, and scene editing.

3.3 Scene Decomposition and Role Assignment

Object Parsing. First, we detect all objects in the scene using Florence-2 [35], a multitask VLM that operates 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) [36] to compute precise segmentation masks, along with confidence scores and labels for each object.

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 Florence-2, 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) of the image height or width. All other remaining objects are considered as potential candidates for editing.

3.4 Scene Editing

For each candidate distractor object, NICE performs one of three edit operations on the copies of the original images (see Figure 2). 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

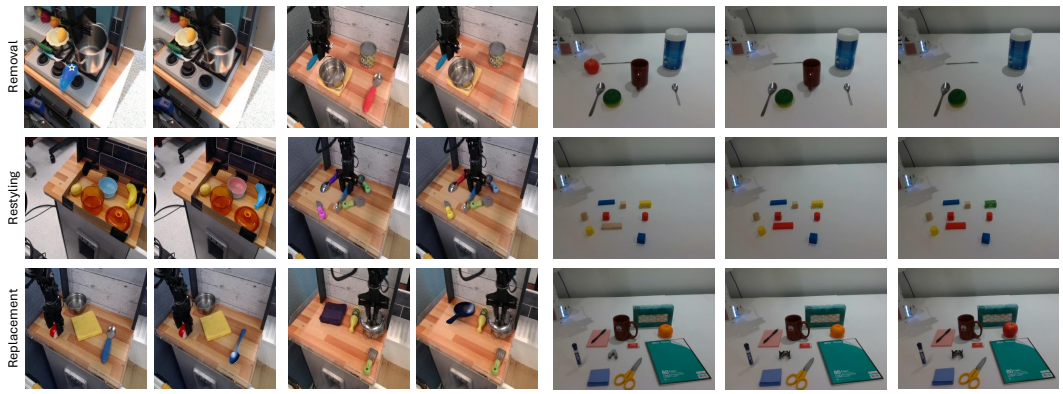

Figure 2: Examples of data enhancement a) using our pipeline on Bridge [37] and b) real-world replication used for evaluation of our pipeline.

target). This mask is then dilated with a hyperparameter dil to smooth the edges and cover the original object's shadow. Finally, we remove the combined distractor mask and apply LaMa [38] to fill the region 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) [39], 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 empirically. 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 a semantically similar one. More specifically, after masking out the target region of the image along with dilation, we use the Stable Diffusion inpainting model [40] 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 realistic novel objects.

For replacement, we can employ two different strategies. 1- Generate the object with different features, by passing its name to diffusion model and ask to alter it. 2- Generate a semantically similar yet visually distinct object variant (e.g. replacing a spoon with a different type of spoon as shown in Figure 2a). This allows us to generate a novel scene while maintaining the context. For this purpose, we use Deepseek-r1:7b [41] via the Ollama framework [42] to generate a description of a household object similar in size to the original one, which is then fed into the Stable Diffusion model [40]. In our experiments, we found using such a small language model suffices for accurate prompting in order to generate similar objects.

4 Evaluation

4.1 Background Consistency

A key consideration 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

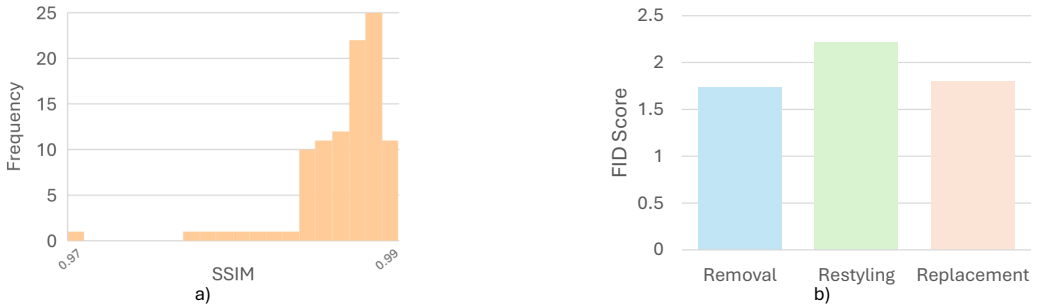

Figure 3: a) distribution of SSIM values for removal operation and b) FID score of the three enhancement strategies on real-world samples.

Table 1: Average prediction accuracy (APA)(%) across different clutter levels using RoboPoint.

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)

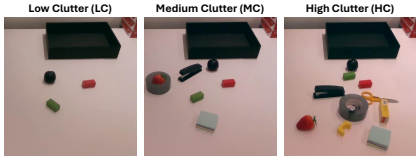

Figure 4: Samples of scenes with different level of clutter.

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 scene by removing one object at a time, for a total of 100 real-world images (see example in Figure 2b). We then replicate these changes using our pipeline and compare to real images using the SSIM metric [43]. As shown in Figure 3a, our method generally yields a very high score on generated samples, indicating its accuracy in reconstructing the background.

4.2 Data Generation Realism

Following the similar procedure as in 4.1, 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). Samples of real-world data are shown in Figure 2b. Using our pipeline, we then replicate the scene alterations and compute Fréchet Inception Distance (FID) [44] between the generated and real-world captured images. As shown in Figure 3b, 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.

4.3 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 [45], a state-of-the-art vision-language-model that predicts spatial affordance in free space, which then can be used for any downstream robotic task.

For this experiment, as shown in Figure 4, we consider scenes with three levels of clutter: *low clutter* with 1-2 objects, *medium clutter* with 5-8 objects, and *high clutter* 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 scenes, the objects are densely placed to increase difficulty. Following the protocol in [45], we report the results using *average prediction accuracy (APA)*, which measures the percentage of predicted points that fall within the ground-truth target mask.

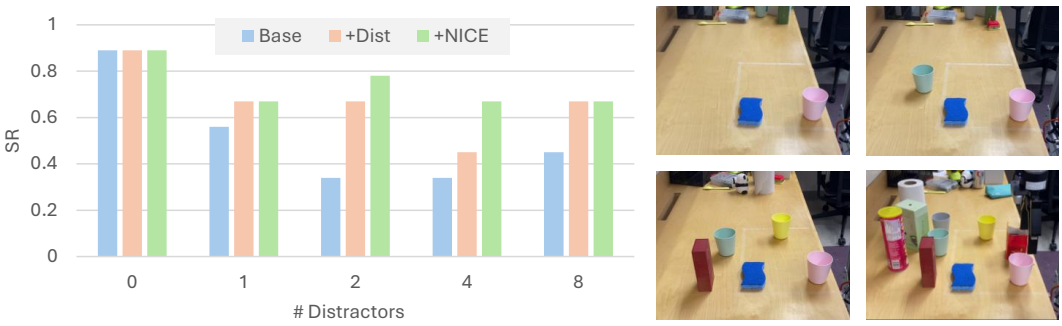

Figure 5: (**Left**) Performance of the manipulation policy π_0, finetuned on three data configurations. (**Right**) Example experiment scenes with varying numbers of distractors.

As shown in Table 1 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.

4.4 Robotic Manipulation in Clutter

To evaluate the benefits of our enhancement pipeline on manipulation data, we choose a pick-and-place task involving a pink cup (as shown in Figure 5). In the base scenario, the scene only includes the target. We further populate the scene by adding 1, 2, 4, and 8 distractors to the environment. For each setup, we vary the target object's position across 9 uniformly spaced locations within the operating area and repeat the experiment.

For manipulation, we employ a VLA policy, π_0 [46], pre-trained on Open X-Embodiment [47]. We finetune three variants of π_0 with different data configurations: (1) Base, containing only the target object; (2) +Dist, which adds real data with 9 variations of the target object in the presence of 8 distractors; and (3) +NICE, which incorporates context-enhanced data generated from training samples with distractors. As shown in Figure 5, using data that includes distractors improves the success rate (SR) on most cluttered scenes. However, by enriching data using NICE, we can further boost the performance by 11% on scenes with 2 distractors and 22% on scenes with 4 distractors.

Table 2: Average performance of the policy using different datasets. Direction of arrows shows higher or lower values are better.

Dataset	SR ↑	CR↓	oCR↓
Base	0.51	0.38	0.15
+Dist	0.65	0.09	0.07
+NICE	**0.74**	**0.06**	**0.02**

Besides improvement in success rate, our approach can potentially improve safety. To highlight this effect, we report on total collision rate (CR), involving contacts with any non-target objects. We also report on obstacle collision rate (oCR), involving collisions with objects other than currently targeted one whether it is the intended target object or not. i.e. in oCR we exclude contacts due to target confusion. According to Table 2, our method not only results in best average SR (+23% compared to base and +9% compared to Dist), it also lowers collision by 3% and 5% on CR and oCR, respectively.

5 Conclusion

In this work, we presented 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.

6 Limitations

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.

Due to the limited scope of the paper, we only examined the impact of our framework on spatial affordance prediction and a pick-and-place manipulation task. It is reasonable to assume that visual confusion or operational confinement caused by distractors can have different degree 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 as well.

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, T.-H. Wang, Y. Wang, K. Fragkiadaki, Z. Erickson, D. Held, and C. Gan. 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*, 86(5): 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*, 5(5):172331, 2018.

[11] M. A. Petilli, F. Marini, and R. Daini. Distractor context manipulation in visual search: How expectations modulate proactive control. *Cognition*, 196:104129, 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*, 78(1-2):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*, 29:135–147, 2019.

[14] Y. Li, H. Zhu, Y. Cheng, W. Wang, C. S. Teo, C. Xiang, P. Vadakkepat, and T. H. Lee. 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] N. Di Palo and E. Johns. Keypoint action tokens enable in-context imitation learning in robotics. In *RSS*, 2024.

[22] H. Kim, Y. Ohmura, and Y. Kuniyoshi. Using human gaze to improve robustness against irrelevant objects in robot manipulation tasks. *RAL*, 5(3):4415–4422, 2020.

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

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

[25] 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*, 110 (2):62, 2024.

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

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

[28] A. Xie, L. Lee, T. Xiao, and C. Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation. In *ICRA*, 2024.

[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] 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.

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

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

[37] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine. Bridgedata v2: A dataset for robot learning at scale. In *CoRL*, 2023.

[38] 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.

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

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

[41] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[42] 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*, pages 23–35. 2025.

[43] A. Hore and D. Ziou. Image quality metrics: Psnr vs. ssim. In *ICPR*, 2010.

[44] 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.

[45] 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.

[46] 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.

[47] 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.