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

Sim2real Image Translation Enables Viewpoint-Robust Policies from Fixed-Camera Datasets

Jeremiah Coholich[1], Justin Wit[1], Robert Azarcon[1], Zsolt Kira[1]

***Abstract*—Vision-based policies for robot manipulation have achieved significant recent success, but are still brittle to distribution shifts such as camera viewpoint variations. Robot demonstration data is scarce and often lacks appropriate variation in camera viewpoints. Simulation offers a way to collect robot demonstrations at scale with comprehensive coverage of different viewpoints, but presents a visual sim2real challenge. To bridge this gap, we propose MANGO – an unpaired image translation method with a novel segmentation-conditioned InfoNCE loss, a highly-regularized discriminator design, and a modified PatchNCE loss. We find that these elements are crucial for maintaining viewpoint consistency during sim2real translation. When training MANGO, we only require a small amount of fixed-camera data from the real world, but show that our method can generate diverse unseen viewpoints by translating simulated observations. In this setting, MANGO outperforms all other image translation methods we tested. In certain real-world tabletop manipulation tasks, MANGO augmentation increases shifted-view success rates by over 40 percentage points compared to policies trained without augmentation. For more results, visit: https://www.jeremiahcoholich.com/mango.**

I. Introduction

Significant progress has been made in developing vision-based imitation-learning policies for robot manipulation. Performant single-task architectures [1]–[4] and intuitive teleoperation methods [1], [5], [6] have given way to large robot datasets and multi-task Vision-Language-Action models (VLAs) [7]–[11]. However, the robot demonstrations used to train these models are scarce and labor-intensive to generate. In comparison to web-scraped vision-language datasets, robot learning datasets often lack diversity which results in models with poor zero-shot performance to new setups.

For example, many tabletop manipulation datasets employ fixed, third person cameras for observations [3], [4], [12]–[16], making downstream policies brittle to camera viewpoint shifts. Indeed, we have observed that when robot polices are trained on fixed-camera datasets, changes in camera viewpoint during deployment cause success rates to crater (Table IV). Cameras are often fixed during demo collection to ensure consistent visual observations, avoid repeated calibration with depth or motion capture sensors, or simply for convenience. Generalizing to truly unseen viewpoints is difficult because a change in viewpoint affects the entire scene, and the model must implicitly estimate the robot's position relative to the new camera position.

[1]Institute of Robotics and Intelligent Machine, Georgia Institute of Technology, Atlanta, GA, USA. Emails: {jcoholich, jwit3, razarcon3, zkira}@gatech.edu.

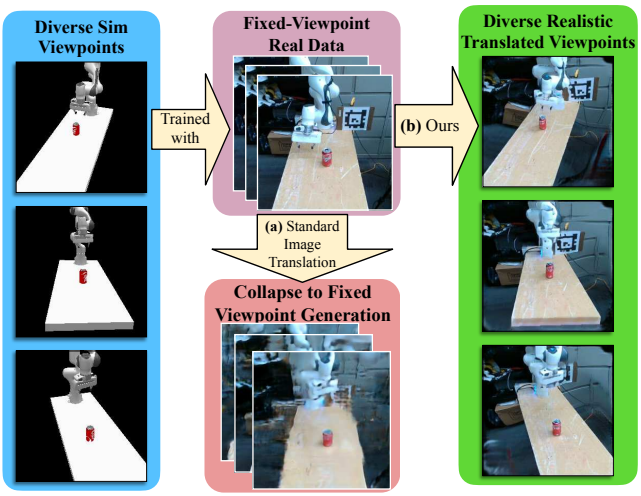

Fig. 1: (a) Standard image translation methods fail to generalize to new viewpoints when trained on a fixed-viewpoint target domain dataset. (b) Our method, MANGO, enables realistic generation of unseen viewpoints, which are used to improve the robustness of downstream robot polices.

To augment fixed-camera robot data, we propose to collect simulated demonstrations from a simple digital twin via task and motion planning whose visual observations are taken from diverse camera viewpoints. Crucially, we train a novel image translation model for bridging the visual sim2real gap. We name our proposed approach Multiview Augmentation with Novel Generated Observations, or MANGO. With MANGO, we simultaneously bypass manual data collection and solve the viewpoint diversity issue. Our MANGO image translation model is trained on a small real-world dataset collected from a single fixed camera plus a larger simulation dataset of images with segmentations. After training, MANGO is able to translate diverse simulation viewpoints to unseen real-world viewpoints. While our method relies on a digital twin, we use a deliberately simple simulation rendered with low-fidelity OpenGL settings, without extensive visual engineering. The entire pipeline yields synthetic demonstrations with realistic and varied camera viewpoints.

We argue for GAN-based models over diffusion models for robot data augmentation. Robot demonstration datasets contain many image observations which are downsampled to small image sizes. A robot dataset for a small, single-task policy may contain upwards of 180,000 images, assuming

979-8-3315-8160-2/26/$31.00 ©2026 IEEE

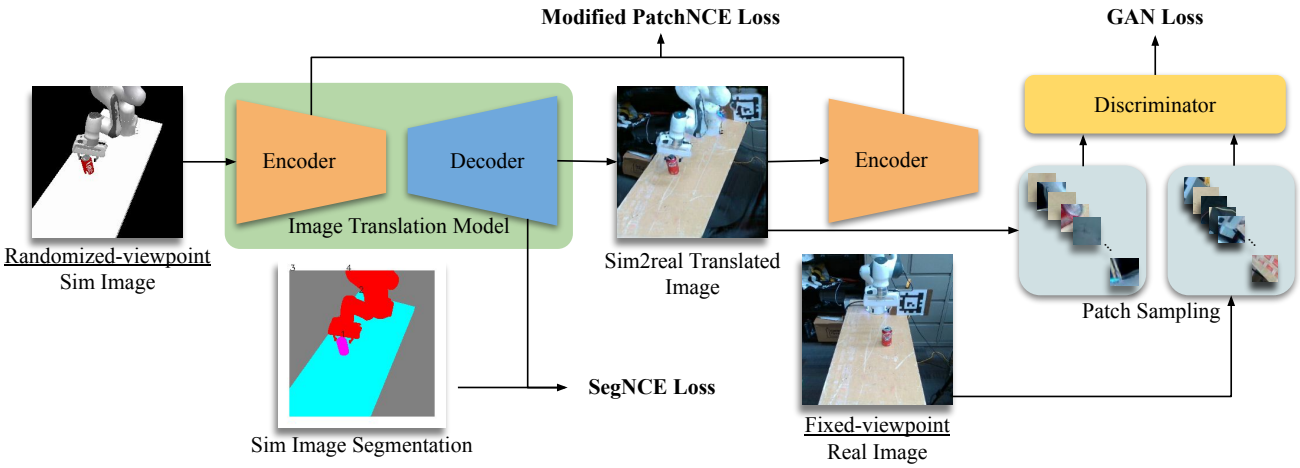

Fig. 2: MANGO is trained on unpaired real and sim images, specifically a real dataset obtained with a fixed camera and a simulation dataset with diverse camera viewpoints. To ensure the simulation viewpoint is preserved during translation, we employ a novel segmentation-based InfoNCE loss, a modified PatchNCE loss from [17], and a random patch sampling and rotation process to regularize the discriminator D.

150 demonstrations collected at 20 Hz. Using ZeroNVS [18], a state-of-the-art diffusion model for novel viewpoint generation, augmenting this dataset would take 435 GPU-hours. Training an ACT policy on such a dataset would only require ~ 5 hours on a single GPU, meaning that diffusion model-based augmentation would severely bottleneck compute requirements. Our lightweight GAN-based approach is approximately 2,700x times faster than ZeroNVS, making it practical for visual dataset augmentation.

Our core contributions are as follows:

1) MANGO: A sim2real image translation method incorporating a novel, segmentation-informed contrastive InfoNCE loss capable of preserving unseen simulation viewpoints during translation
2) Experimental proof that demonstrations generated with MANGO improves downstream robot policy robustness to shifts in camera position, even in comparison to a diffusion-based method
3) Analysis of why our method is successful on the domain of robot demonstration datasets in comparison to the many other approaches developed for the generic problem of unpaired image translation

II. Related Work

A. Visual Sim2Real Translation for Robotics

Simulation is a popular tool among roboticists, but to actually leverage simulated data for real robot policies requires bridging the sim2real gap. For visual observations, one option is to improve simulator realism [19], however this is a labor-intensive engineering effort which must be done for every scene. Domain-randomization of lighting, colors, and textures is another option [20]–[22], but determining the degree and types of randomizations to apply is a challenge, and policies trained with domain randomization often sacrifice performance for robustness.

Image-to-image translation has been proposed to cross this sim2real gap by learning from data. A wide variety of unpaired image-to-image translation architectures exist [17], [23]–[28]. To translate from sim2real, roboticists can directly train these models on datasets of image observations collected from simulation and the real world. For example, in [29] and [30], the authors train an unmodified CycleGAN to translate visual observations for grasping and navigation.

Others have tailored these methods to incorporate specific knowledge about the robot and downstream application. DigitalTwin-CycleGAN adds an action cycle-consistency loss to CycleGAN for a sim2real visual grasping task [31]. This loss makes the image translation model dependent on learning a successful grasping policy concurrently. RL-CycleGAN incorporates Q-function consistency on translated images [32], where the Q-function is obtained while learning a task-specific RL policy. RetinaGAN enforces cycle consistency with an object detector which requires thousands of labeled images to train beforehand [33]. GraspGAN trains an image translation model without cycle-consistency and instead enforces accurate image content translation through a grasp success predictor [34]. Additionally, they include an auxiliary generator objective of reproducing the ground-truth sim image segmentation. CyCADA unifies these methods under a general "task loss" framework [35]. In contrast, MANGO is agnostic to the downstream learning algorithm, enabling us to train one image translation model for many tasks.

Diffusion models [36] have emerged as the primary architecture for image generation over generative adversarial networks (GANs) [37], with some exceptions [38]–[40]. However, we find that for the specific domain of unpaired

image-to-image translation, GANs obtain results competitive with the best diffusion approaches [24]. We hypothesize that that the output multimodality of diffusion models is a disadvantage when the style and content of the generated image are tightly-specified by the input image and target domain dataset, respectively. MANGO uses the GAN loss; however in theory our novel segmentation-based InfoNCE loss could be applied to any image translation architecture containing a generator network with spatially-indexed latent feature maps.

B. Robot Policy Viewpoint Invariance

RoboNet offered early proof that training a robot policy on multiple views enables generalization to viewpoint shifts [41]. Multi-view Masked World Models (MV-MWM) [42] demonstrates impressive robustness to camera viewpoints by training a viewpoint-invariant visual encoder and task-specific world model. MoVie [43] achieves view generalization by adapting an image encoder to the novel views during test-time. In contrast, MANGO does not require any test-time adaptation or real-world images from viewpoints outside of the fixed-camera images used for training. [44] trains an RL policy that is robust to single-camera viewpoint changes after learning from a teacher policy trained with a multi-view observation. VISTA leverages pretrained models with 3D priors to generate novel viewpoints given a single real-world image observation [45]. However, since they do not use simulation-generated demonstrations they are unable to generate new robot trajectories and must rely on human demonstration collection. Additionally, VISTA finetunes the ZeroNVS [18] diffusion model, which suffers from high resource requirements as discussed in Section I.

Learned 3D representations are inherently viewpoint invariant in theory, but still overfit to the specific 2D sensor locations in practice. Additionally, building strong 3D implicit or explicit representations typically requires more data than a single RGB sensor. For example, GROOT [46] achieves impressive viewpoint invariance but requires task-specific object annotations. 3D Diffusion Policy and 3D Diffusor-Actor both build 3D scene representations from calibrated RGBD cameras, but are shown to be brittle to the viewpoints used for this synthesis [3], [16]. Adapt3R achieves greater viewpoint robustness through only mapping embedding vectors to 3D instead creating entire scene pointclouds, but still requires multiple calibrated RGBD external sensors [15]. MANGO demonstrations viewpoint robustness with only a single external RGB camera.

III. Method

A. Image-to-image Translation

We propose MANGO, a novel unpaired image-to-image translation method to translate visual observations from sim to real (Figure 2). MANGO is trained on a small, fixed-viewpoint dataset of real images yet is capable of accurately translating viewpoint-diverse observations from a simple digital twin to realistic unseen viewpoints.

The objective of unpaired image-to-image translation is to translate images from domain A to domain B without access to a paired dataset of images $\mathcal{D}_{paired} = \{d_A,\ d_B | d_A \in A,\ d_b \in B\}_{i=0}^{N}$. Instead, we learn from two disjoint datasets $\mathcal{D}_A$ and $\mathcal{D}_B$, where our domain A is simulation, our domain B is the real-world, and $|\mathcal{D}_A| > |\mathcal{D}_B|$. This problem is considered unsupervised because there is no label, or ground-truth image, in $\mathcal{D}_B$ that images in $\mathcal{D}_A$ map to.

Image translators like MANGO must change the style of the input image while preserving its content. We employ the GAN architecture with a highly-regularized discriminator to learn the target domain style. For accurate content preservation, we use the InfoNCE [47] loss between input and output image features in a similar style as CUT [17], but with a modified scoring function. Additionally, we propose a novel segmentation-based InfoNCE loss on generator features.

B. Style Loss

We use the standard GAN loss [37] to enforce target domain style on the generated images, given by Equation 1. G is the generator network, and D is the discriminator network.

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G, D, \mathcal{D}_A, \mathcal{D}_B) = \mathbb{E}_{x\sim\mathcal{D}_B} \log(D(x)) \\ +\ \mathbb{E}_{y\sim\mathcal{D}_A} \log(1 - D(G(y)))\end{aligned} \tag{1}$$

One assumption underlying image-translation GANs, such as CycleGAN [23] and CUT [17], is that the shared attributes among all images in $\mathcal{D}_B$ constitute the target domain "style". However, our real-world robot image observations in $\mathcal{D}_B$ only differ from one another in robot and object poses. Much of the image content, such as the background and tabletop, is nearly identical in all images in $\mathcal{D}_B$. A naïve discriminator will memorize the repetitive details and force the generator to recreate them. To mitigate this problem, Pix2pix [48] proposed a "PatchGAN", where the discriminator only receives local image patches and cannot therefore enforce global image elements. We take this a step further and randomly sample patch locations and apply per-patch random rotations. This process is shown in Figure 2. The result is a highly-regularized discriminator capable of enforcing the style of $\mathcal{D}_B$ on images with viewpoints not seen in $\mathcal{D}_B$.

C. Content Loss

MANGO content translation losses consists of a modified version of the PatchNCE loss [17] and a novel segmentation-based NCE loss.

1) Modified PatchNCE Loss: The PatchNCE loss consists of an InfoNCE loss across encoder features generated by a source domain image $d_A \in D_A$ and its corresponding translated output image $G(d_A)$. For an input image d, we randomly sample N latent features from the encoder at L different layers. We call the set of features at layer l $\mathcal{Z}_l$ and $|\mathcal{Z}_l| = N\ \forall\ l \in \{l_0, ..., l_L\}$. The translated image $\hat{d}$ is passed through the encoder again to obtain $\hat{\mathcal{Z}}_l\ \forall\ l \in \{l_0, ..., l_L\}$. All $\hat{\mathcal{Z}}_l$ are obtained from the same feature map indices as in $\mathcal{Z}_l$.

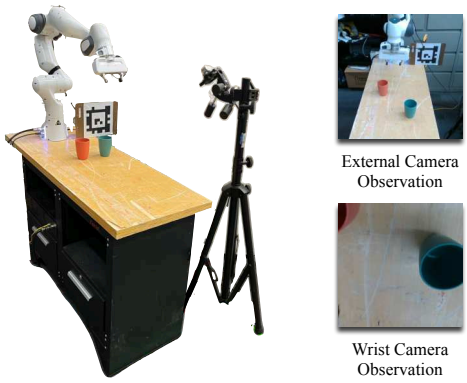

Fig. 3: Our real-world robot setup, including a Franka Emika Panda arm, a wrist camera, and an external camera that is only repositioned for evaluations.

The InfoNCE loss for feature i in encoder layer l is given by Equation 2. This is the categorical cross-entropy loss on the probability that a feature $\mathbf{z} \in \mathcal{Z}$ will be correctly classified as the corresponding feature in $\hat{\mathcal{Z}}$, based on a scoring function $\rho_l(\cdot)$. τ is a temperature hyperparameter.

$$\ell_{\text{NCE}}(l, \mathbf{z}, \hat{\mathcal{Z}}, i) = -\log \left[\frac{\exp\left(\rho_l(\mathbf{z}, \hat{\mathbf{z}}_i)/\tau\right)}{\sum_{\hat{\mathbf{z}} \in \hat{\mathcal{Z}}} \exp\left(\rho_l(\mathbf{z}, \hat{\mathbf{z}})/\tau\right)} \right] \tag{2}$$

$\rho_l(\cdot)$ is defined in Equation 3. Features $\mathbf{z}_i$ and $\mathbf{z}_j$ are passed through a function H_l then scored with cosine similarity.

$$\rho_l(\mathbf{z}_i, \mathbf{z}_j) = \frac{H_l(\mathbf{z}_i) \cdot H_l(\mathbf{z}_j)}{\|H_l(\mathbf{z}_i)\| \|H_l(\mathbf{z}_j)\|} \tag{3}$$

The full loss is given in Equation 4

$$\mathcal{L}_{\text{PatchNCE}}(G, H, \mathcal{D}) = \mathbb{E}_{d \sim \mathcal{D}} \sum_{l=1}^{L} \sum_{i=1}^{|\mathcal{Z}_l|} \ell_{\text{NCE}}(l,\ \mathbf{z}_{l,i},\ \hat{\mathcal{Z}}_l,\ i) \tag{4}$$

The reasoning behind Equation 2 is that input and output features from the same feature map locations are "positive" samples and should have high similarity scores. All other features are "negative" samples and should be repelled. However, we observe that many different input image patches are highly similar due to repeated patterns or textures in robot image datasets, which include background elements, the tabletop, etc. Furthermore, in the simulated image dataset $\mathcal{D}_A$, these regions all have *identical* pixel values due to simplistic rendering. Therefore, Equation 2 will repel many false negative features.

To mitigate this, we train MANGO with a modified scoring function. If the cosine similarity of a negative sample exceeds a threshold θ, we multiply the score by a factor $0 \leq \alpha < 1$. The modified scoring function is given by Equation 5.

$$\tilde{\rho}_l(\mathbf{z}_i, \mathbf{z}_j) = \begin{cases} \alpha\, \rho_l(\mathbf{z}_i, \mathbf{z}_j) & \text{if } \rho_l(\mathbf{z}_i, \mathbf{z}_j) > \theta \text{ and } i \neq j, \\ \rho_l(\mathbf{z}_i, \mathbf{z}_j) & \text{otherwise.} \end{cases} \tag{5}$$

We have empirically found this to be more effective than increasing τ. We denote the modified NCE loss which uses the scoring function in Equation 5 as $\tilde{\mathcal{L}}_{\text{PatchNCE}}$. In practice, we set $\alpha = 0.5$ and $\theta = 0.9$.

2) Segmentation NCE Loss: We leverage the simulator used to generate $\mathcal{D}_A$ to obtain ground-truth segmentation maps for each image. We propose an InfoNCE loss which clusters generator features by segmentation category in order to ensure that object boundaries are preserved during translation. We note that for object-centric manipulation, this is a crucial aspect to preserve.

Each image in $\mathcal{D}_A$ contains C segmentation classes and each feature $\mathbf{z}_i \in \mathcal{Z}$ generated from $d \sim D_A$ has an associated class label $y_i \in \mathcal{Y}$. In the case when a layer's feature map is of a lower resolution than the input image, we scale the image segmentation with nearest-neighbors downsampling to obtain y_i.

The Segmentation NCE (SegNCE) loss is defined in Equation 6. In contrast to ℓ_{NCE} as shown in Equation 2, there are multiple positive samples for the query feature $\mathbf{z}_i$. All features that are in the same segmentation class as the query feature are classified as positive samples and indexed by j.

In Equation 2, the target distribution for the cross-entropy loss is a one-hot vector. In Equation 6 the target distribution is a uniform distribution over features from the same segmentation class and zero elsewhere.

Here, we use the original scoring function $\rho_l(\cdot)$ defined in Equation 3; since we are operating with ground-truth image segmentations, there are no false negatives.

$$\ell_{\text{SegNCE}}(l, \mathcal{Z}, i, \mathcal{Y}) = \frac{1}{\left|\left\{ j \,\middle|\, \substack{j \in 1..|\mathcal{Z}| \\ y_j = y_i \\ i \neq j} \right\}\right|} \sum_{\left\{ j \,\middle|\, \substack{j \in 1..|\mathcal{Z}| \\ y_j = y_i \\ i \neq j} \right\}} \ell_{\text{NCE}}(l, \mathbf{z}_i, \mathcal{Z}, j) \tag{6}$$

The full loss term is given in Equation 7.

$$\mathcal{L}_{\text{SegNCE}}(G, H, \mathcal{D}_A) = \mathbb{E}_{d \sim \mathcal{D}_A} \sum_{l=1}^{L} \sum_{i=1}^{S} \ell_{\text{SegNCE}}(l, \mathcal{Z}_l, i, \mathcal{Y}_l) \tag{7}$$

The SegNCE loss is computed from input image generator features only.

D. Model training

The total loss function for G is given in equation 8. We include an identity PatchNCE loss for regularization following [17]. The full discriminator loss is given in Equation 9 and is simply the GAN objective.

$$\begin{aligned} \mathcal{L}_G = \tilde{\mathcal{L}}_{\text{PatchNCE}}(G, H, \mathcal{D}_A) \\ + \tilde{\mathcal{L}}_{\text{PatchNCE}}(G, H, \mathcal{D}_B) \\ + \mathcal{L}_{\text{SegNCE}}(G, H, \mathcal{D}_A) \\ + \mathcal{L}_{\text{GAN}}(G, D, \mathcal{D}_A, \mathcal{D}_B) \end{aligned} \tag{8}$$

$$\mathcal{L}_D = -\mathcal{L}_{\text{GAN}}(G, D, \mathcal{D}_A, \mathcal{D}_B) \tag{9}$$

IV. EXPERIMENTS

Our experiments are designed to answer the following questions:

1) How well can MANGO translate sim images to unseen unseen real-world viewpoints?
2) Are imitation-learning policies trained with synthetic data from MANGO more robust to shifts in camera position?
3) How does MANGO compare to baselines such as domain randomization and diffusion-based image augmentation methods?

A. Image Translation with MANGO

1) Training Details: Our generator G is a 12M parameter ResNet-based network. The discriminator D is a wider three-layer CNN with 11M parameters. Additionally, we parameterize the H_l in Equations 3 and 5 as a two-layer MLP with 700k parameters.

We first benchmark the image translation method on observations from a "`pick up coke`" task. $\mathcal{D}_A$ is a dataset of 8,098 image observations from simulation with camera viewpoints randomized within a box of dimensions (100, 100, 84) cm (L×W×H). $\mathcal{D}_B$ contains 3,094 images obtained from roughly 10 minutes of real-world teleoperated play data. All images are cropped and scaled to 256x256. Training the image translation model takes approximately 20 hours on a single RTX 2080 Ti GPU.

We curate three test datasets: Fixed View, Randomized View, and Wrist View. Each test set contains 128 sim/real image pairs. The Fixed View test set contains sim and real images from the same fixed view as $\mathcal{D}_B$. The randomized view testset contains both sim and real viewpoints taken from a camera placed randomly within a box of dimensions (100, 100, 84) cm. In order to create paired images for the Randomized View testset, we use the robot state in conjunction with AprilTags for coke can and camera pose estimation. Note that these are not needed for the deployment of MANGO, only for test set creation.

TABLE I: Sim2real Unpaired Image Translation FID Scores. Each score is averaged from two models trained with different seeds. D is the discriminator and $\mathcal{D}_A$ is the sim training image dataset.

Method	Fixed View FID(↓)	Randomized View FID(↓)	Wrist View FID(↓)
No Translation	340.3	297.4	268.8
CUT [17]	412.3	373.9	266.3
CycleGAN [23]	393.9	359.9	265.8
Basic D	371.3	318.5	293.5
Without SegNCE loss	267.0	207.5	199.9
Without $\tilde{\rho}_l(\cdot)$	192.9	184.1	195.3
Fixed-cam $\mathcal{D}_A$	**108.0**	198.7	202.5
MANGO	182.3	**160.9**	**191.3**

2) Results: Table I gives the scores of our proposed method against baselines and ablations. MANGO obtains the lowest FID score by 23 points on the Randomized Camera testset. The patch discriminator D has the largest impact on the FID score for all testsets. Translated image examples are given in Figure 5.

TABLE III: Average pairwise LPIPS on natural image datasets. A core challenge in robot learning is lack of dataset diversity as compared to web data.
† Results computed on 10^3 randomly sampled images.

Dataset	Average Pariwise LPIPS (↑)
Laion-5B†	0.725
ImageNet†	**0.819**
Cifar-10†	0.221
Cifar-100†	0.250
Horse2zebra $\mathcal{D}_A$ (Horses)	0.747
Horse2zebra $\mathcal{D}_B$ (Zebras)	0.765
Seg2Cityscapes $\mathcal{D}_B$ (Real)	0.548
`pick up coke` $\mathcal{D}_B$ (Real)	0.155

Note that while the relative FID scores correlate well with relative image quality, the scores in Table I are high compared to the numbers reported in other literature. We posit that this is due to the small size of our test sets and that our robotics lab scene may be out-of-distribution for the Inception network used for FID Score.

We hypothesize that off-the-shelf methods like CUT and CycleGAN struggle with robotics datasets due to their lack of diversity. Typically, unpaired image-to-image translation methods are benchmarked on computer vision datasets containing diverse images scraped from the internet. In comparison, robotics datasets may be collected from a single setup. To support this claim, we compute the average pairwise LPIPS (a learned perceptual distance metric) [49] on various image datasets as a measure of diversity. As shown in Table III, our $\mathcal{D}_B$ shows the lowest score.

B. Simulation Experiments

We evaluate MANGO on simulated tasks from Robomimic [50] and Mimicgen [51] and give the results in Table II. Specifically, we measure the FID scores of data generated with MANGO, as well as the success rates of behavioral-cloning policies trained with the data. Instead of sim2real, we create two visually disparate simulation environments and run sim2sim experiments. Example observations from each domain and MANGO translations are given in Figure 4. We benchmark MANGO against policies trained on single-camera observations, untranslated Domain A data, and VISTA [45]. To provide a fair comparison, we only train MANGO on image observations from the same camera viewpoints and tasks used by VISTA, and report metrics on the same six evaluation tasks. "Unseen object" tasks contain objects not seen in training, "shared object" contains objects seen during training but in different contexts, and "cross-embodiment" tasks are seen performed by the Rethink Sawyer robot instead of Franka Panda. We train MANGO with fixed camera data from the six tasks used for testing in domain B, and varied camera data of the eight training tasks in domain A. This ensures our model does not see views

TABLE II: Sim2sim Experiment Results. Success rates are the average of 50 rollouts. FID scores are computed across target-domain generated images and target-domain oracle images.

	Unseen Object				Shared Object				Cross-Embodiment			
Data Augmentation	**Threading**↑	*FID*↓	**Hammer**↑	*FID*↓	**Coffee**↑	*FID*↓	**Stack**↑	*FID*↓	**PickPlace**↑	*FID*↓	**Nut Asm.**↑	*FID*↓
None (Fixed Camera Only)	10.67	*54.46*	18.00	*59.45*	14.00	*43.83*	47.33	***25.86***	30.67	*86.28*	13.33	*45.13*
Depth est.+Repoj.	2.67	*80.47*	20.67	*61.44*	9.33	*68.77*	42.67	*61.29*	31.33	*93.92*	10.00	*61.23*
VISTA	28.00	***48.19***	56.00	*44.95*	40.67	*43.99*	66.67	*38.60*	**45.33**	***64.57***	28.67	***25.96***
Untranslated (Domain B)	1.33	*74.56*	0.00	*58.50*	0.00	*48.61*	2.67	*76.5*	44.67	*80.27*	0.00	*99.59*
MANGO (Domain B → A)	**30.00**	*50.03*	**86.00**	***32.11***	**64.67**	***42.45***	**71.33**	*37.47*	13.33	*100.12*	**45.33**	*53.88*
Simulator (Oracle)	57.33	–	100.00	–	80.67	–	86.67	–	86.00	–	64.00	–

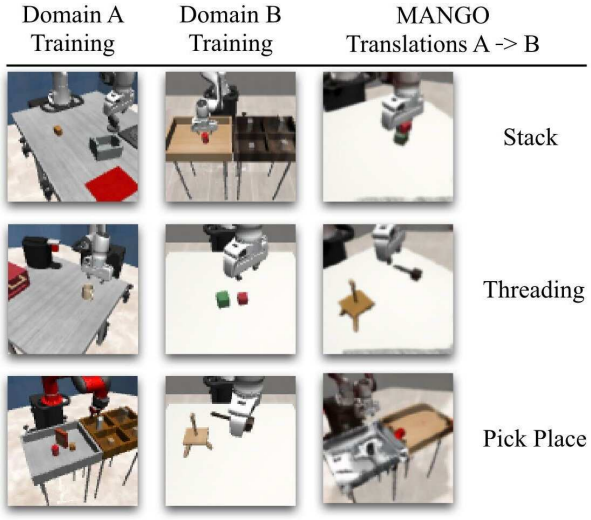

Fig. 4: Sim2Sim training data with translations by MANGO for three of the tasks included in Table II

from the test viewpoint distribution, aside from the single fixed view. MANGO-trained BC policies obtain the highest success rate on 5 out of 6 tasks.

C. Real Robot Experiments

We benchmark MANGO on four real-world robotic manipulation tasks: "`pick up coke`", "`stack cups`", "`close laptop`", and "`stack blocks`". We train a single image translation model for all tasks, where $\mathcal{D}_A$ and $\mathcal{D}_B$ consist of 59,520 simulated observations and 35,294 fixed-camera real observations from tasks. Our real robot setup is shown in Figure 3.

1) Policy Training and Evaluation Details: We train action chunking transformer (ACT) [1] policies on synthetic data generated by MANGO. The generated data is translated from our digital twin demonstrations which leverage the task and motion planner from RLBench [52]. All policies are cotrained with 150 human-teleoperated demos with fixed-camera and wrist camera observations. ACT was chosen to isolate the effects of our generated image data as it does not incorporate any pretraining or language conditioning. We train each ACT policy for 10k epochs with a chunk size of 20. Rollouts are done without temporal aggregation.

We compare ACT models trained on MANGO data to strong sim2real and viewpoint augmentation baselines, depicted in Figure 5. For domain randomization we follow [20] and randomize color, texture, and lighting for all objects in the scene. VISTA is a viewpoint augmentation method that leverages a fine-tuned ZeroNVS model [45]. Unlike MANGO, VISTA has built-in rejection sampling for generated images based on LPIPS distance to the original images.

We evaluate each trained policy on 10 variations per task on four different camera viewpoints. The first camera viewpoint is the fixed-camera which the real demonstrations and MANGO training data are collected from. The three shifted viewpoints, depicted in Figure 5, are aggregated into the "Shifted Cam" column in Table IV.

2) Results: Results for MANGO-trained policies and baselines are given in Table IV. With MANGO, we observe large increases in for viewpoint robustness as compared to models trained on fixed-camera human demonstrations only. We also observe that our method is necessary to bridge the sim2real gap, since the policies trained on sim demos without translation perform consistently worse than MANGO.

TABLE IV: Success rates for imitation learning policies across tasks, viewpoint shifts, and data augmentation methods. The only method comparable to MANGO is VISTA, which uses a 4.5B parameter pretrained model in contrast MANGO's 35M parameters.

Data Augmentation	**Pick up coke**		**Stack cups**		**Close laptop**		**Stack blocks**	
	Fixed Cam	**Shifted Cams**	**Fixed Cam**	**Shifted Cams**	**Fixed Cam**	**Shifted Cams**	**Fixed Cam**	**Shifted Cams**
None	**8/10**	5/30	8/10	1/30	10/10	19/30	8/10	1/30
Sim	6/10	14/30	**9/10**	5/30	10/10	20/30	**9/10**	9/30
Sim DR	**8/10**	19/30	8/10	5/30	10/10	25/30	7/10	3/30
VISTA	7/10	**23/30**	7/10	8/30	10/10	**29/30**	8/10	**18/30**
Ours	6/10	17/30	8/10	**13/30**	10/10	22/30	**9/10**	11/30

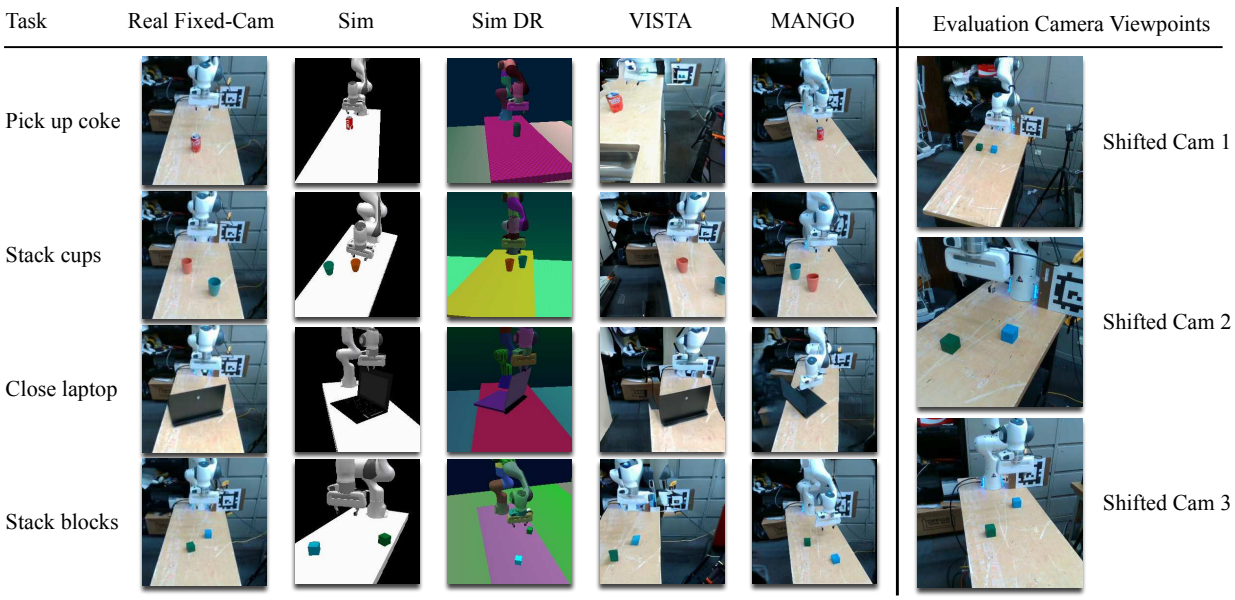

Fig. 5: **Left:** Sample image observations for each task and data-augmentation method from Table IV. **Right:** The three shifted viewpoints that comprise our "Shifted Cams" evaluations in Table IV

V. Discussion and Conclusion

We propose a novel image generation method, MANGO, which is trained on simulated and real robot data. We observe that with only fixed-camera real data, our novel SegNCE loss, discriminator design, and modified PatchNCE loss enable translation of unseen-viewpoint simulated observations to realistic real-world observations. MANGO-generated data improves the robustness of downstream imitation learning policies to camera shift, as demonstrated by greatly increased success rates six simulated and four real-world manipulation tasks. We observe that our method is superior for sim2real translation in this setting, beating all other image translation methods we tested.

There are several limitations to this work. MANGO still requires a small amount of real-world data from the evaluation domain for training. Additionally, we struggle to beat VISTA in 3 out of 4 real world tasks when evaluated on the shifted camera viewpoints. The primary benefit of MANGO is its ability to preserve image quality during translations with a lightweight model that is practical for translating robot demonstration datasets with hundreds of thousands of image observations. MANGO, including data generation and image translation, requires less than 0.2% of the GPU-hours required by VISTA. However, larger pretrained models inherit a more general understanding of 3D geometry and scenes. Incorporating MANGO's novel loss formulations, specifically the segmentation-based InfoNCE loss, into heavier pretrained models for sim2real visual observation translation or real2real augmentation is a promising direction for future work.

References

[1] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[2] L. Wang, X. Chen, J. Zhao, and K. He, "Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers," *Advances in neural information processing systems*, vol. 37, pp. 124 420–124 450, 2024.

[3] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," *arXiv preprint arXiv:2402.10885*, 2024.

[4] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[5] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto, "Open teach: A versatile teleoperation system for robotic manipulation," *arXiv preprint arXiv:2403.07870*, 2024.

[6] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 12 156–12 163.

[7] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[8] A. Szot, B. Mazoure, H. Agrawal, R. D. Hjelm, Z. Kira, and A. Toshev, "Grounding multimodal large language models in actions," *Advances in Neural Information Processing Systems*, vol. 37, pp. 20 198–20 224, 2025.

[9] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "Pi_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[10] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, "Pi_0.5: A vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[11] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[12] S. Haldar, V. Mathur, D. Yarats, and L. Pinto, "Watch and match: Supercharging imitation with regularized optimal transport," in *Conference on Robot Learning*. PMLR, 2023, pp. 32–43.

[13] 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 *Conference on Robot Learning (CoRL)*, 2023.

[14] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *2019 International conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 8943–8950.

[15] A. Wilcox, M. Ghanem, M. Moghani, P. Barroso, B. Joffe, and A. Garg, "Adapt3r: Adaptive 3d scene representation for domain transfer in imitation learning," *arXiv preprint arXiv:2503.04877*, 2025.

[16] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," *arXiv preprint arXiv:2403.03954*, 2024.

[17] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 319–345.

[18] K. Sargent, Z. Li, T. Shah, C. Herrmann, H.-X. Yu, Y. Zhang, E. R. Chan, D. Lagun, L. Fei-Fei, D. Sun *et al.*, "Zeronvs: Zero-shot 360-degree view synthesis from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9420–9429.

[19] X. Li, K. Hsu, J. Gu, K. Pertsch, O. Mees, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani *et al.*, "Evaluating real-world robot manipulation policies in simulation," *arXiv preprint arXiv:2405.05941*, 2024.

[20] L. Pinto, M. Andrychowicz, P. Welinder, W. Zaremba, and P. Abbeel, "Asymmetric actor critic for image-based robot learning," *arXiv preprint arXiv:1710.06542*, 2017.

[21] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray *et al.*, "Learning dexterous in-hand manipulation," *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020.

[22] R. Garcia, R. Strudel, S. Chen, E. Arlaud, I. Laptev, and C. Schmid, "Robust visual sim-to-real transfer for robotic manipulation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. ieee, 2023, pp. 992–999.

[23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[24] M. Zhao, F. Bao, C. Li, and J. Zhu, "Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3609–3623, 2022.

[25] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.

[26] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2849–2857.

[27] V. Sushko, E. Schönfeld, D. Zhang, J. Gall, B. Schiele, and A. Khoreva, "You only need adversarial supervision for semantic image synthesis," *arXiv preprint arXiv:2012.04781*, 2020.

[28] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "Ilvr: Conditioning method for denoising diffusion probabilistic models," *arXiv preprint arXiv:2108.02938*, 2021.

[29] H. Zhang, H. Liang, L. Cong, J. Lyu, L. Zeng, P. Feng, and J. Zhang, "Reinforcement learning based pushing and grasping objects from ungraspable poses," *arXiv preprint arXiv:2302.13328*, 2023.

[30] J. Truong, S. Chernova, and D. Batra, "Bi-directional domain adaptation for sim2real transfer of embodied navigation agents," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2634–2641, 2021.

[31] D. Liu, Y. Chen, and Z. Wu, "Digital twin (dt)-cyclegan: Enabling zero-shot sim-to-real transfer of visual grasping models," *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 2421–2428, 2023.

[32] K. Rao, C. Harris, A. Irpan, S. Levine, J. Ibarz, and M. Khansari, "Rl-cyclegan: Reinforcement learning aware simulation-to-real," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 157–11 166.

[33] D. Ho, K. Rao, Z. Xu, E. Jang, M. Khansari, and Y. Bai, "Retinagan: An object-aware approach to sim-to-real transfer," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 10 920–10 926.

[34] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige *et al.*, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 4243–4250.

[35] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. Pmlr, 2018, pp. 1989–1998.

[36] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[37] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[38] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," *arXiv preprint arXiv:2206.10789*, vol. 2, no. 3, p. 5, 2022.

[39] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park, "Scaling up gans for text-to-image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 124–10 134.

[40] A. Sauer, T. Karras, S. Laine, A. Geiger, and T. Aila, "Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis," in *International conference on machine learning*. PMLR, 2023, pp. 30 105–30 118.

[41] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn, "Robonet: Large-scale multi-robot learning," *arXiv preprint arXiv:1910.11215*, 2019.

[42] Y. Seo, J. Kim, S. James, K. Lee, J. Shin, and P. Abbeel, "Multi-view masked world models for visual robotic manipulation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 30 613–30 632.

[43] S. Yang, Y. Ze, and H. Xu, "Movie: Visual model-based policy adaptation for view generalization," *Advances in Neural Information Processing Systems*, vol. 36, pp. 21 507–21 523, 2023.

[44] C. Acar, K. Binici, A. Tekirdağ, and Y. Wu, "Visual-policy learning through multi-camera view to single-camera view knowledge distillation for robot manipulation tasks," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 691–698, 2023.

[45] S. Tian, B. Wulfe, K. Sargent, K. Liu, S. Zakharov, V. Guizilini, and J. Wu, "View-invariant policy learning via zero-shot novel view synthesis," *arXiv preprint arXiv:2409.03685*, 2024.

[46] Y. Zhu, Z. Jiang, P. Stone, and Y. Zhu, "Learning generalizable manipulation policies with object-centric 3d representations," *arXiv preprint arXiv:2310.14386*, 2023.

[47] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[48] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[49] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[50] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," in *arXiv preprint arXiv:2108.03298*, 2021.

[51] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox, "Mimicgen: A data generation system for scalable robot learning using human demonstrations," *arXiv preprint arXiv:2310.17596*, 2023.

[52] S. James, Z. Ma, D. Rovick Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, 2020.