Benchmarking View-Invariance: Comparing Visual Representations and Augmentations for Diffusion Policies

Abstract

Visuomotor policies, which map visual observations to robot actions, have advanced significantly with the emergence of diffusion-based models. However, these policies often struggle to generalize to changes in visual conditions like unseen objects, background clutter, and camera viewpoint changes. This work specifically investigates whether diffusion policies can be made more robust to viewpoint changes through different visual representations, such as 3D point clouds or pre-trained latent embeddings. An ideal solution would be view-invariant, modularly integrate into existing policy training pipelines, and only require minimal camera setups. We systematically compare different representations and existing data-augmentation strategies, such as View Synthesis Augmentation (VISTA). To quantify robustness, we introduce a benchmark with three different levels of increasingly difficult viewpoint shifts. Our experiments demonstrate that policies trained on explicit 3D representations maintain higher performance across all perturbation levels compared to alternative representations and augmentation-based baselines.

1. Introduction

Learning robust visuomotor policies that map visual observations directly to robot actions has been a long-standing goal in robotics. Recently, diffusion-based generative models [1] have emerged as a powerful framework for imitation learning, demonstrating a superior ability to handle multi-modal action distributions and high-dimensional action spaces compared to traditional behavior cloning. However, despite their success in controlled environments, these policies often exhibit a significant "generalization gap" when deployed under varying visual conditions [12]. Among the various factors of environmental variation, changes in camera viewpoints remain one of the most challenging obstacles to reliable deployment, often causing failure in policies trained on single-view datasets [10, 3].

In prior work, two approaches to mitigate viewpoint dependency have been data augmentation and invariant representations. Augmentation-based strategies like VISTA [11] and SPARTN [14], leverage novel-view synthesis (NVS) via diffusion models or Neural Radiance Fields (NeRF) [7] to synthetically expand the training distribution. While effective, these methods are often computationally expensive and can be limited by the generative quality of the underlying NVS model. Alternatively, representation-based methods seek to extract features that are inherently robust to geometric shifts, ranging from pre-trained latent embeddings from vision foundation models [6] to unified view representations [4].

One promising direction is the use of explicit 3D representations. Recent works like 3D Diffusion Policy (DP3) [13] and 3D Diffuser Actor [2] have demonstrated that point clouds can provide a more geometrically grounded signal for policy learning. By operating in 3D space, these models can theoretically achieve better view-invariance than their 2D counterparts. However, there remains a lack of systematic experiments under severe viewpoint perturbations to compare these explicit 3D representations to implicit latent embeddings or data augmentation.

In this work, we investigate the robustness of diffusion policies to viewpoint changes through a systematic comparison of visual representations. We evaluate explicit 3D point clouds, pre-trained latent embeddings, and standard 2D inputs augmented with the VISTA framework. To rigorously quantify performance, we introduce a new benchmark featuring three levels of viewpoint shifts, ranging from minor perturbations to significant displacements. Our results indicate that explicit 3D representations provide the most consistent performance across all levels of difficulty, offering a modular and view-invariant solution for robust robotic manipulation.

The main contributions of this report are summarized as follows:

- We conduct a systematic comparative study of 3D point clouds, latent embeddings, and NVS-based data augmentation (VISTA) for diffusion-based policy learning.
- We introduce a benchmark with three, increasingly difficult levels of viewpoint shifts to evaluate policy generalization in robotic manipulation.

- We demonstrate that explicit 3D representations outperform alternative visual representations and augmentation baselines in maintaining success rates under out-of-distribution camera poses.

2. Related Works

Diffusion Policies for Visuomotor Control The introduction of Diffusion Policy [1] has shifted the paradigm of imitation learning from simple behavior cloning to conditional generative modeling. By representing the action distribution as a denoising diffusion process, these models effectively handle multi-modality and high-dimensional action spaces that traditionally pose challenges for regression-based methods. While Diffusion Policy demonstrates significant robustness to visual distractions and task variations, its performance remains sensitive to the quality of the underlying visual representation and the breadth of the training data. This has led to a growing interest in enhancing these policies through things like better visual representations and data augmentation strategies.

Visual Foundation Models and Pre-trained Encoders Pretrained, frozen visual encoders trained on large datasets can be used to improve the generalization of robot policies.

- **R3M** [8] leverages egocentric human videos (Ego4D) via time-contrastive learning and video-language alignment to capture features relevant to object manipulation.
- **VC-1** [5] scales this approach further by training on over 4,000 hours of video and ImageNet data using Masked Auto-Encoding (MAE), aiming to create a general-purpose "visual cortex" for embodied AI.
- **DINO** [9] utilize self-distillation with Vision Transformers (ViT) to learn dense, semantic features that have shown remarkable zero-shot performance in tasks such as depth estimation and object grounding [**?**].

While these encoders provide rich semantic information, they are primarily 2D-centric. In this work, we contrast these implicit latent representations against explicit 3D point clouds to determine which is more resilient to perturbations common in real-world deployment.

View-Invariant Policy Learning Achieving robustness to camera viewpoint changes is a critical hurdle for generalizable manipulation. Standard practices often rely on multi-view setups [3] or simple 2D augmentations (e.g., cropping, color jittering), which typically fail under significant geometric shifts. Recently, View Synthesis Augmentation (VISTA) [11] proposed leveraging large-scale novel view synthesis (NVS) models, such as ZeroNVS, to synthetically generate diverse camera perspectives from single-view demonstrations. While VISTA provides a zero-shot path toward viewpoint invariance by augmenting the 2D image distribution, our work explores whether explicit 3D representations can provide a more computationally efficient and modular alternative for achieving similar robustness.

3D Representations in Robot Learning Explicitly incorporating 3D information into the policy pipeline has recently shown promising results. Methods like 3D Diffusion Policy (DP3) [13] and 3D Diffuser Actor [2] move away from 2D pixel-grids toward point clouds or 3D scene embeddings. By processing geometry directly, these models can inherently account for spatial relationships.

3. Methodology

In this section, we detail the simulation environments, model architectures, training setups, and the evaluation setup used to quantify the view invariance of policies.

3.1. Simulation Environment and Tasks

Our experiments are conducted in the RoboSuite and RoboMimic simulation frameworks. We evaluate our method on three distinct manipulation tasks:

- **Lift**: A single-object lifting task with a maximum of 100 steps.
- **Stack**: A block stacking task (*stack*) requiring 150 steps for completion.

3.2. Observation and Action Spaces

The observation space O consists of multi-modal inputs. Visual observations include a 128×128 RGB images from an exocentric camera and an egocentric, wrist-mounted camera. Proprioceptive data includes the 3D end-effector position, 4D orientation quaternion, and a 2D gripper state. For explicit 3D models, point clouds are extracted using image and depth information and processed using a PointNet encoder.

The action space $A \in \mathbb{R}^{10}$ is represented using absolute coordinates: 3D position, 6D rotation representation, and a 1D gripper command. All datasets are stored in HDF5 format with absolute action space trajectories.

3.3. Policy Architecture

We utilize the Diffusion Policy framework with a UNet-based denoising backbone. The diffusion process employs a Denoising Diffusion Probabilistic Model (DDPM) scheduler with $T = 100$ timesteps. The UNet architecture features down-sampling dimensions of $[512, 1024, 2048]$ and a diffusion step embedding of 128.

3.4. Approaches

We evaluate the view invariance of 3 approaches :

1. **Pretrained Encoders:** Visual encoders trained on large, non-robotic datasets. Both camera views are passed in separately and the latent representations are inputted into the policy.
    - ResNet
    - DINOV2
    - DINOV3
    - R3M
2. **Point clouds:** These are generated using the image, depth, camera extrinsics and intrinsics data. The point cloud is then passed into the PointNet encoder and the output is passed into the policy.
3. **VISTA:** We use the "Oracle" approach from the VISTA paper where instead of generating new views using NVS models, multiple ground truth views are recorded for an action trajectory and at each timestep, one of them is randomly chosen as the exocentric view. The egocentric view remains the same. This increases the size of the dataset, and aims to make the policy invariant to changes in the camera view. We use two versions of this approach, "easy" and "hard". The dataset used in these approaches contains views from the respective evaluation distribution as detailed in section 3.6. This gives the approach a large advantage since it is the only approach that sees views similar to its evaluation during training.

3.5. Training Hyperparameters

All policies are trained using the diffusion policy codebase. Policies are trained for 200 epochs using a batch size of 64. We employ the AdamW optimizer ($\beta = [0.95, 0.999]$, weight decay $= 10^{-6}$) and an Exponential Moving Average (EMA) with $\gamma = 0.75$. The learning rate is set to 10^{-4} with a cosine decay schedule following a 500-step linear warmup. During training, images are randomly cropped to 112×112 and normalized using ImageNet statistics. The policy uses a horizon of 16 steps, with an observation history of 2 steps and an execution horizon of 8 steps.

3.6. Evaluation

To evaluate generalization to camera viewpoint shifts, we define a robustness benchmark categorized into three difficulty levels based on the values of five different parameters that measure different dimensions of displacement from the training view:

1. θ: Rotation around world Z-axis
2. ϕ: Rotation around camera's X-axis
3. ψ: Rotation around camera's Z-axis
4. d: Distance from the origin in the XY plane
5. h: Vertical offset from default height

For each difficulty level, value ranges can be defined for each of these parameters. When values are sampled from these ranges, we get the parameters required to perturb the default viewpoint according to the set difficulty. This leads to the three levels:

Parameter	Easy	Medium	Hard
θ (Yaw)	$\pm 10^\circ$	$\pm[10, 25]^\circ$	$\pm[25, 45]^\circ$
ϕ (Pitch)	0°	$\pm 0.5^\circ$	$\pm[0.5, 1]^\circ$
ψ (Roll)	$\pm 5^\circ$	$\pm[5, 7.5]^\circ$	$\pm[7.5, 10]^\circ$
d (Distance)	0	± 0.05	$\pm[0.05, 0.1]$
h (Height)	0	± 0.05	$\pm[0.05, 0.1]$

where easy represents minimal angular perturbations, medium represents moderate angular and translational shifts and hard represents significant out-of-distribution viewpoint shifts. Visual comparisons of these levels can be found in the Appendix.

Evaluation is performed on 20 different environments. For each model, the top-2 checkpoints (selected by by highest success rate) are evaluated across 5 views per difficulty level. Final results are reported as the average of these checkpoints using two primary metrics:

- **Success Rate**: Percentage of episodes where the entire trask is completed.
- **Mean Score**: Average of the maximum rewards achieved across episodes.

3.7. Training Hardware and Times

Models were trained using a single NVIDIA RTX 3090 GPU. Training runs typically required 8 hours, while the multi-view evaluation runs required 3 hours.

4. Results

We analyze performance using two primary metrics. **Success Rate (SR):** Percentage of times where task was done in its entirety. **Completion (C):** How close it got to finishing the task

4.1. Comparing Visual Encoders

We first evaluate the performance of various 2D visual encoders. As shown in Table 1, DINOv3 demonstrates the best performance across all three difficulty levels on the Lift Task. While R3M shows high performance in the Easy setting, its performance drops sharply in the Medium and Hard settings. For the more complex Stack Two task (Table 2) shows that all 2D representations struggle significantly as

the difficulty increases. They are able to get great performance with an unperturbed camera view, their performance is significantly worse for signifcant view changes. While DINOv3 maintains the highest SR in the Easy condition (59.0%), success rates for all 2D encoders fall toward zero in the Hard setting, indicating a lack of inherent geometric understanding required for non-trivial tasks under viewpoint shifts.

Method	No Change	Easy	Medium	Hard
ResNet	100%	74.0 / 0.83	11.0 / 0.32	10.0 / 0.33
R3M	100%	**92.5 / 0.95**	18.5 / 0.39	22.0 / 0.42
DINOv2	100%	48.0 / 0.65	53.0 / 0.69	52.0 / 0.68
DINOv3	100%	**96.5 / 0.98**	**63.0 / 0.75**	**69.0 / 0.80**

Table 1: Comparison of encoders on *Lift* (SR / C).

Method	No Change	Easy	Medium	Hard
ResNet	100%	45.0 / 0.56	2.0 / 0.13	1.0 / 0.10
R3M	85%	33.5 / 0.54	0.0 / 0.05	0.0 / 0.11
DINOv2	100%	37.5 / 0.51	1.5 / 0.15	1.0 / 0.18
DINOv3	100%	**59.0 / 0.74**	**7.0 / 0.21**	0.0 / 0.09

Table 2: Comparison of encoders on *Stack* (SR / C).

4.2. Comparing Approaches

We further compare the best-performing 2D encoder (DINOv3) against explicit 3D Point Cloud representations and View Synthesis Augmentation (Oracle) strategies outlined in the methods section. In the Lift task (Table 3), both the Oracle (Hard) and the Point Cloud approach achieve near-perfect generalization across all perturbation levels, reaching 100% success rates even in the *Hard* setting. This result is more impressive for the Point Cloud method since similar views are seen during training for the Oracle approach. Both approaches significantly outperform the best 2D foundation models.

In the Stack task (Table 4), DinoV3 is unable to generalize to medium and hard settings. The Oracle (Easy) approach—which represents an ideal augmentation for small shifts—performs the best on the easy setting but fails to generalize to harder evals. The Oracle (Hard) approach has the best performance for the hard setting since it has seen similar views during training, but it is outperformed by the Point Cloud approach on the easy and medium settings. These results suggest that while augmentation can improve local robustness, explicit 3D representations provide superior global view-invariance for robotic manipulation.

5. Conclusion

In this work, we systematically investigated the robustness of diffusion-based visuomotor policies to significant viewpoint shifts. By establishing a rigorous evaluation benchmark across three levels of difficulty, we compared the efficacy of 2D pretrained foundation models, data augmentation strategies (VISTA Oracle), and explicit 3D point cloud representations.

Method	No Change	Easy	Medium	Hard
DINOv3	100%	96.5 / 0.98	63.0 / 0.75	69.0 / 0.80
Oracle (Easy)	100%	**100 / 1.00**	51.0 / 0.66	49.0 / 0.65
Oracle (Hard)	100%	**100 / 1.00**	**100 / 1.00**	**100 / 1.00**
Point Cloud	100%	**100 / 1.00**	**99.5 / 1.00**	**100 / 1.00**

Table 3: Robustness comparison on *Lift* (SR / C).

Method	No Change	Easy	Medium	Hard
DINOv3	100%	59.0 / 0.74	7.0 / 0.21	0.0 / 0.09
Oracle (Easy)	95%	**96.0 / 0.97**	21.0 / 0.40	8.5 / 0.31
Oracle (Hard)	95%	**83.0 / 0.93**	**62.0 / 0.80**	**90.0 / 0.96**
Point Cloud	100%	**88.0 / 0.93**	**74.5 / 0.85**	**69.0 / 0.79**

Table 4: Robustness comparison on *Stack* (SR / C).

Our results demonstrate that while 2D visual foundation models—particularly DINOv3—exhibit some resilience to minimal perturbations, they largely fail to generalize to moderate or significant viewpoint shifts in non-trivial tasks like block stacking. Data augmentation via the VISTA Oracle approach provides substantial local robustness; however, it remains constrained by the distribution of views encountered during training. Even when provided with ground-truth views from the target evaluation distribution, augmentation strategies were often outperformed by geometric representations in settings they had not explicitly seen but could theoretically interpolate.

Conversely, explicit 3D point cloud representations demonstrated superior global view-invariance. The point cloud approach achieved near-perfect generalization on the Lift task and maintained high success rates in the Hard setting of the Stack task, despite not having seen perturbed views during the training phase. These findings highlight that explicitly modeling the 3D geometry of the scene provides a more robust and modular pathway for generalizable robot learning than relying on 2D perceptual priors or expansive data augmentation. Future work will explore the integration of these 3D representations into real-world systems where depth data may be noisier than the simulated environments used in this study.

References

[1] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023. 1, 2

[2] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. In *Conference on Robot Learning (CoRL)*, 2024. 1, 2

[3] Litian Liang, Liuyu Bian, Caiwei Xiao, Jialin Zhang, Linghao Chen, Isabella Liu, Fanbo Xiang, Zhiao Huang, and Hao Su. Robo360: a 3d omnispective multi-material robotic manipulation dataset. *arXiv preprint arXiv:2312.06686*, 2023. 1, 2

[4] Fanfan Liu, Feng Yan, Liming Zheng, Chengjian Feng, Yiyang Huang, and Lin Ma. Robouniview: Visual-language model with unified view representation for robotic manipulation. *arXiv preprint arXiv:2406.18977*, 2024. 1

[5] Arjun Majumdar, Karaman Yadav, Sergio Arnaud, Yecheng Jason Ma, Lawrence Chen, Sneha Silwal, Aryan Jain, Vincent Amancharla, Manasi Shukla, Somya Gunapati, et al. Where are we in out-of-distribution visual representation learning for control? In *Conference on Robot Learning (CoRL)*, 2023. 2

[6] Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liang-Yan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. In *Advances in Neural Information Processing Systems*, 2024. 1

[7] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[8] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2022. 2

[9] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Mehdi SM Sajjadi, Xingyuan Liao, Adeline Vazquez, Yuki Asano, Josh Tobin, Mathilde Gazeau, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[10] Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation. In *Robotics: Science and Systems (RSS)*, 2024. 1

[11] Stephen Tian, Blake Wulfe, Kyle Sargent, Katherine Liu, Sergey Zakharov, Vitor Guizilini, and Jiajun Wu. View-invariant policy learning via zero-shot novel view synthesis. In *Conference on Robot Learning (CoRL)*, 2024. 1, 2

[12] Annie Xie, Lisa Lee, Ted Xiao, and Chelsea Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation. In *Conference on Robot Learning (CoRL)*, 2023. 1

[13] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Robotics: Science and Systems (RSS)*, 2024. 1, 2

[14] Allan Zhou, Moo Jin Kim, Lirui Wang, Pete Florence, and Chelsea Finn. Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis. In *Conference on Robot Learning (CoRL)*, 2023. 1

6. Appendix

This section visualizes the exocentric camera views used in our study. Figure 1 shows the static viewpoint used during training, while Figure 2 illustrates the three difficulty levels used to benchmark view-invariance.

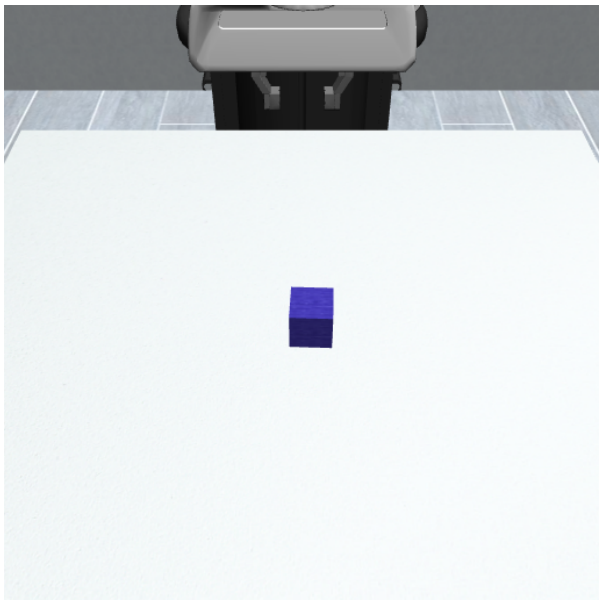
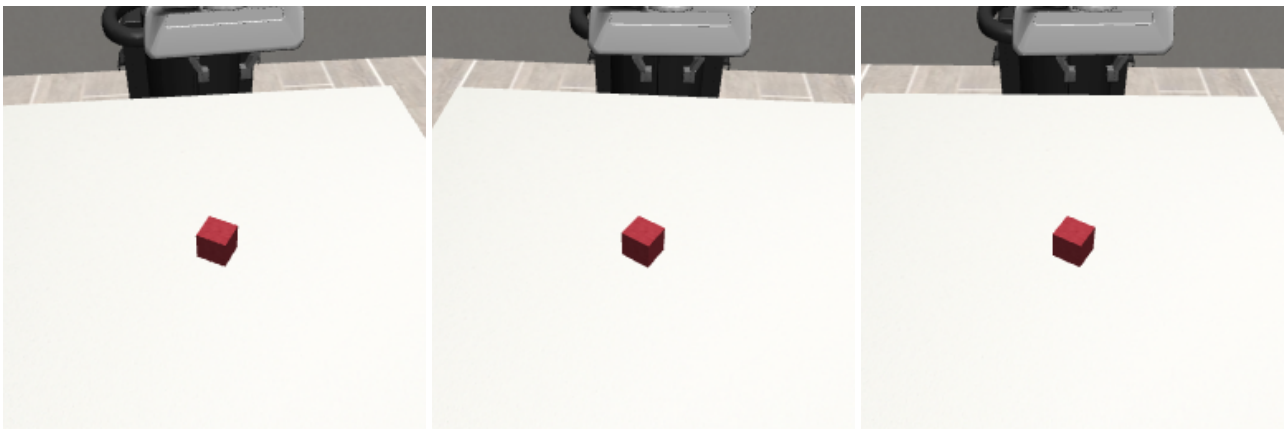
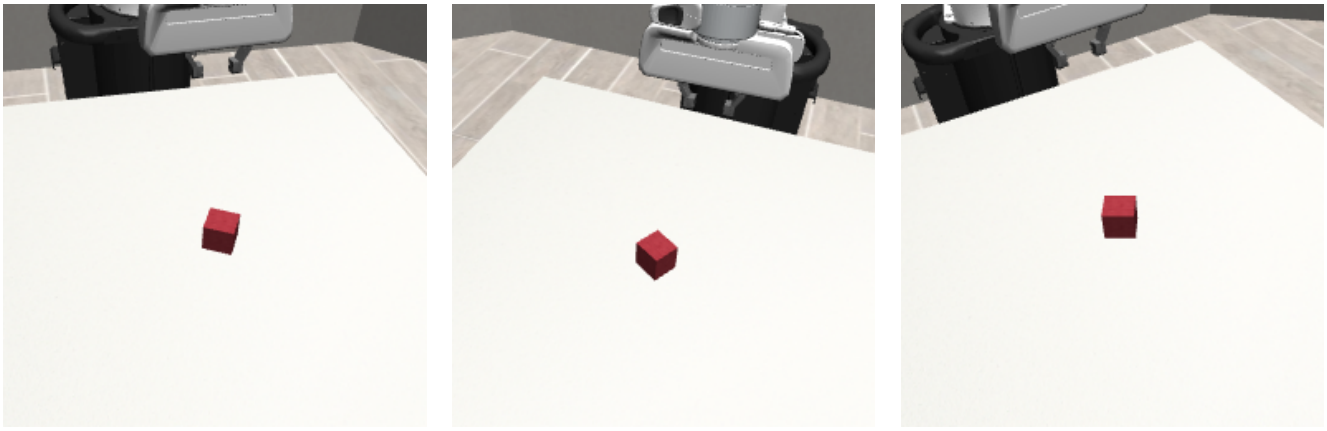


Figure 1: Default Training View: The exocentric camera remains at this fixed pose during all expert demonstration collection.



Easy Viewpoint Samples



Medium Viewpoint Samples

Hard Viewpoint Samples

Figure 2: Visual breakdown of the evaluation benchmark. Each row displays three random samples from the respective difficulty range. The **Easy** row shows minor perturbations. The **Medium** row adds translational offsets and some harder perturbations. The **Hard** row represents significant out-of-distribution shifts.