DisDP: Robust Imitation Learning via Disentangled Diffusion Policies

Pankhuri Vanjani[1], Paul Mattes[1], Kevin Daniel Kuryshev[1], Xiaogang Jia[1], Vedant Dave[2] and Rudolf Lioutikov[1]
[1]Intuitive Robots Lab, Karlsruhe Institute of Technology, Germany [2]Montanuniversität Leoben

***Abstract*—This work introduces Disentangled Diffusion Policy (DisDP), an Imitation Learning (IL) method that enhances robustness. Robot policies have to be robust against different perturbations, including sensor noise, complete sensor dropout and environmental variations. Existing IL methods struggle to generalize under such conditions, as they typically assume consistent, noise-free inputs. To address this limitation, DisDP structures sensors into shared and private representations, preserving global features while retaining details from individual sensors. Additionally, Disentangled Behavior Cloning (DisBC) is introduced, a disentangled Behavior Cloning (BC) policy, to demonstrate the general applicance of disentanglement for IL. This structured representation improves resilience against sensor dropouts and perturbations. Evaluations on The Colosseum and Libero benchmarks demonstrate that disentangled policies achieve better performance in general and exhibit greater robustness to perturbations compared to their baseline policies.**

I. Introduction

For robots to be deployed on a large scale across various applications, they have to be robust against different perturbations, including environmental variations, sensor noise, and complete sensor modality dropout. Sensor modality dropout refers to the unavailability of sensors during inference, that have been available during training. While current research has explored environmental variations, and perturbations in behavior learning scenarios [31], sensor modality dropout remains an understudied challenge. Methods that attempt to address this issue often fail to generalize on complex, multi-view robotic benchmarks [37, 9], exposing a critical vulnerability in the safety and reliability of current IL-based policies. To tackle this challenge, we propose **Disentangled Diffusion Policy (DisDP)**, a method that disentangles the latent space of different sensor modalities into shared and private embeddings.

Integrating multiple sensors improves policy robustness, especially when individual inputs are noisy or unreliable [20, 36, 38, 24, 16]. However, such setups face challenges like calibration errors, noise, and sensor failures. Most existing methods assume reliable inputs at inference [36, 38, 34], limiting robustness to sensor dropout, as shown in Section IV. This work focuses on vision-based IL policies and their resilience to missing or noisy camera inputs (Figure 1). Imitation Learning (IL) [1, 29] is widely used for acquiring complex behaviors, with recent multi-task IL methods achieving strong results across diverse tasks [30, 35, 36, 9, 34, 32, 4, 6].

Despite these advancements, most IL methods rely on latent representations not explicitly designed to handle noisy or missing sensory data, making them vulnerable to sensor degradation or dropout. This paper introduces disentangled representations for IL policies to enhance robustness and interpretability. By separating sensor inputs into shared and private embeddings, our approach addresses key challenges such as camera noise, sensor dropout, and environmental perturbations. We apply this disentanglement to both a score-based diffusion policy [32, 34] and a Transformer-based BC model [26]. Extensive evaluations show that this structure improves overall performance and significantly reduces degradation under unreliable sensing conditions.

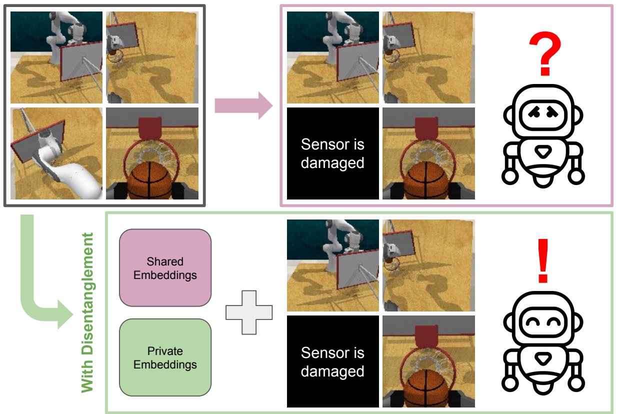

Fig. 1: Robotic policies depend on multiple sensory inputs, making them susceptible to sensor failures. This work investigates how disentangling sensory information into shared and private embeddings can enable robust policy learning under sensor dropouts.

II. Related Work

A. Robustness in Behavior Learning

Behavior learning suffers from generalization limitations, often leading to sharp performance degradation in unfamiliar environments due to overfitting and limited adaptability to unseen variations [45, 5, 15, 18, 44]. To address this, various methods have been proposed to improve generalization and robustness under modality dropout [30, 43, 23, 41, 10, 2]. One line of work focuses on estimating missing modalities: SMIL [25] uses Bayesian meta-learning with variational inference to infer the posterior of missing inputs, while CCM [19] applies self-supervision to identify and discard corrupted sensor inputs before reconstructing multimodal representations. However, these methods do not address complete sensory failure.

Several approaches improve robustness by handling missing or irrelevant modalities. Masking-based methods drop or suppress modalities during training based on task relevance [37, 9], while MIL [8] applies masking before policy construction but does not address sensor failures. Hierarchical methods like Nexus [40] and MUSE [39] learn shared and private embeddings via dropout-based training; Nexus averages features, limiting expressiveness, whereas MUSE uses a Product-of-Experts for better integration. In contrast, DisDP uses contrastive learning, avoiding hierarchical structures. Multi-camera RL setups also apply multi-view disentanglement for robustness with partial views [7]. DisDP extends this to imitation learning, evaluated on Colosseum and Libero.

B. Multi View Disentanglement

Multi-view disentanglement aims to separate information into distinct representations, typically decomposing features into shared and private components across views or modalities. Orthogonal denoising autoencoders enforce independence via orthogonality constraints [42], while self-supervised methods minimize inter-view overlap to extract view-specific features [11, 17]. These approaches often combine alignment, orthogonalization, and reconstruction losses to preserve essential shared information. Information-theoretic methods, such as FactorCL [21], further enhance disentanglement by optimizing mutual information bounds, improving generalization by isolating task-relevant features and suppressing irrelevant ones.

In DisDP, disentanglement techniques discussed above are extended to the **multi-task IL** setting, specifically within **diffusion policy frameworks**. The approach is designed to handle complex robotic manipulation tasks, with experiments conducted on diverse benchmarks. The experiments evaluate effectiveness under various sensor conditions.

III. Method

In this work, we address robustness in multi-task IL with multiple input modalities. Robots are trained to imitate expert demonstrations collected from multiple cameras across diverse manipulation tasks. During deployment, these modalities may become unreliable or unavailable due to occlusion, sensor failure, or noise. Our goal is to develop a framework that can robustly handle partial or degraded inputs under such conditions.

A. Problem Formulation

IL aims to train an agent to perform tasks by learning from expert demonstrations. Given a dataset of expert trajectories $\mathcal{D}_{\boldsymbol{\tau}} = \{\boldsymbol{\tau}_i\}_{i=1}^{N}$, where each trajectory

$$\boldsymbol{\tau}_i = ((\boldsymbol{s}_1, \boldsymbol{a}_1), (\boldsymbol{s}_2, \boldsymbol{a}_2), \ldots, (\boldsymbol{s}_K, \boldsymbol{a}_K)) \tag{1}$$

represents a sequence of observed state-action pairs. The objective is to learn a policy $\pi(\boldsymbol{a}|\boldsymbol{s})$ that maps observations $\boldsymbol{s}$ to actions $\boldsymbol{a}$ while minimizing a some distance or divergence to the observed behavior $\mathcal{L}\left(\pi(\boldsymbol{a}|\boldsymbol{s}_k), \boldsymbol{a}_k\right)$. The exact definition of the loss $\mathcal{L}$ depends on the particular IL approach. In a multi-modal IL setting the state information contains multiple modalities, typically across different sensors. In this work these modalities include:

Language instructions $\boldsymbol{L}_k$ provide high-level task annotations. As they are per demonstration, we reuse the same instruction at each timestep: $\boldsymbol{L}_k := \boldsymbol{L}_i$ for $\boldsymbol{s}_k \in \boldsymbol{\tau}_i$.

RGB images $\boldsymbol{I}_k = (I_k^{(1)}, I_k^{(2)}, \ldots, I_k^{(C)})$ capture the scene from C camera viewpoints.

We define **reliability masks** $\boldsymbol{M}_k = (M_k^{(1)}, M_k^{(2)}, \ldots, M_k^{(C)})$ to model sensor noise and availability. $M_k = \mathbf{1}$ indicates fully reliable input, values in $(0, 1)$ denote partial noise, and $M_k = \mathbf{0}$ indicates an unavailable camera. Thus, each state in the framework is defined as

$$\boldsymbol{s}_k = (\boldsymbol{L}_k, \boldsymbol{I}_k \odot \boldsymbol{M}_k) \in \boldsymbol{\mathcal{S}}, \tag{2}$$

with $\odot$ denoting the Hadamard Product and $\boldsymbol{\mathcal{S}}$ denoting the overall state space.

During training, masking is fixed to $M_k = \mathbf{1}$. At inference, it introduces noise or modality dropout based on the evaluation setup. Behavior is not conditioned on raw inputs but on learned embeddings $\boldsymbol{z}_k = \boldsymbol{\phi}(\boldsymbol{s}_k)$, where $\boldsymbol{\phi}$ encodes sensor inputs. Existing approaches either use a single joint embedding [27] or separate embeddings per modality [34, 13, 32, 33, 14]. Theoretically, learning individual embeddings provides mechanisms to improve robustness against modality dropout. In practice, however, the learned policies usually still require the presence of all embeddings and assume reliable information for each. This work in contrast, does not learn embeddings for individual sensors nor single embeddings across all sensors but instead explicitly learns shared embeddings $\boldsymbol{v}$ across sensors and private embeddings $\boldsymbol{u}$ for each sensor

$$\begin{gathered}\left(\boldsymbol{z}_{\boldsymbol{L},k}, \quad \underbrace{\boldsymbol{z}_{I,k}^{(1)},} \quad \underbrace{\boldsymbol{z}_{I,k}^{(2)},} \quad \ldots, \quad \underbrace{\boldsymbol{z}_{I,k}^{(C)}}\right) \Rightarrow \\ \left(\boldsymbol{z}_{\boldsymbol{L},k}, \left(\boldsymbol{v}_{I,k}^{(1)}, \boldsymbol{u}_{I,k}^{(1)}\right), \left(\boldsymbol{v}_{I,k}^{(2)}, \boldsymbol{u}_{I,k}^{(2)}\right), \ldots, \left(\boldsymbol{v}_{I,k}^{(C)}, \boldsymbol{u}_{I,k}^{(C)}\right)\right)\end{gathered} \tag{3}$$

The shared embeddings $\boldsymbol{v}^{(c)}$ contain information that sensor c shares with other sensors, while the private embeddings $\boldsymbol{u}^{(c)}$ contain information that is unique to the sensor. This formulation allows the policy to learn a more robust representation of unreliable sensors. If the sensor c drops out, the private information $\boldsymbol{u}^{(c)}$ of the sensor is not available, however, the information that would have been contained in the shared embedding $\boldsymbol{v}^{(c)}$ is covered by the other sensors.

Finally, recent work introducing action chunking [46] has shown that predicting a sequence of actions generally performs better than generating single step actions. Following this insight the action space is redefined as

$$\bar{\boldsymbol{a}}_k = (\boldsymbol{a}_k, \boldsymbol{a}_{k+1}, \ldots, \boldsymbol{a}_{k+H}) \in \boldsymbol{\mathcal{A}}^H, \tag{4}$$

where H is the prediction horizon, $\boldsymbol{\mathcal{A}}$ denotes the action space and the sequence of actions $(\boldsymbol{a}_k, \boldsymbol{a}_{k+1}, \ldots, \boldsymbol{a}_{k+H})$ was observed in any of the demonstrated trajectories $\boldsymbol{\tau}_i$.

The final policy is represented as $\bar{\boldsymbol{a}}_k \sim \pi\left(\bar{\boldsymbol{a}}_k|\boldsymbol{\phi}(\boldsymbol{s}_k)\right)$ and trained using the dataset

$$\mathcal{D} = \bigcup_{\boldsymbol{\tau}\in\mathcal{D}_{\boldsymbol{\tau}}} \{(\bar{\boldsymbol{a}}, \boldsymbol{s})|(\bar{\boldsymbol{a}}, \boldsymbol{s}) \in \boldsymbol{\tau}\}, \tag{5}$$

which contains pairs of action sequences and states across all demonstrated trajectories. Here, the union $\bigcup$ allows for potentially duplicate entries in the final dataset to maintain the statistical occurrence of state-action pairs.

Multi-view disentanglement: This technique separates representations into shared ($\boldsymbol{v}$) and private ($\boldsymbol{u}$) components across modalities. Shared embeddings capture global, view-consistent features, providing robustness when inputs are missing or noisy. Private embeddings retain fine-grained, view-specific details, improving performance when available.

B. Disentangled Diffusion Policy

Disentangled Diffusion Policy (DisDP) combines a Transformer based encoder-decoder diffusion model [32, 34] with multi-view disentanglement, as illustrated in Figure 2. In the first step, every camera input $I_k^{(c)}$ is embedded using a separate vision encoder. These vision-embeddings are processed through disentanglement branches to obtain a shared embedding $\boldsymbol{v}_k^{(c)}$ and a private embedding $\boldsymbol{u}_k^{(c)}$.

The shared embedding module extracts global features, that are consistent across all camera views $I_k^{(1:C)}$. By focusing on features that remain stable across viewpoints, the shared-embedding encoder provides a robust foundation for downstream tasks, especially when one or more cameras become unreliable, occluded, or noisy. The private embedding module captures fine-grained and view-specific details for each camera view $I_k^{(c)}$. These private features enrich the policy with information unique to each perspective, preserving distinctive cues when global signals are insufficient.

The effective separation of shared and private features is ensured using a contrastive learning approach based on the $\mathrm{InfoNCE}(\boldsymbol{x}, \boldsymbol{x}_+, \boldsymbol{x}_-)$ loss [28, 3]. The contrastive learning loss requires positive $\boldsymbol{x}_+$ and negative samples $\boldsymbol{x}_-$ for each point $\boldsymbol{x}$. The InfoNCE loss then rewards embeddings that are close to positive samples while punishing embeddings that are close to negative samples.

For the shared embedding $\boldsymbol{v}^{(c)}$ of sensor c we obtain the positive samples $\boldsymbol{v}_+^{(c)}$ by sampling shared embeddings of different sensors at the same state. While negative samples $\boldsymbol{v}_-^{(c)}$ are sampled from shared embeddings of different states. The corresponding disentanglement loss is defined as

$$\mathcal{L}_{\text{shared}} = \mathbb{E}_{\boldsymbol{s}\in\mathcal{D}, c\in C, \boldsymbol{v}^{(c)}\in\boldsymbol{\phi}(\boldsymbol{s})}\ \mathrm{InfoNCE}(\boldsymbol{v}^{(c)}, \boldsymbol{v}_+^{(c)}, \boldsymbol{v}_-^{(c)}). \tag{6}$$

For the private embedding $\boldsymbol{u}^{(c)}$ of sensor c the positive samples $\boldsymbol{u}_+^{(c)}$ are drawn form the same camera at different states and the negative samples $\boldsymbol{u}_-^{(c)}$ are drawn from any other sensor at any state. The corresponding disentanglement loss is defined analogously to the shared loss

$$\mathcal{L}_{\text{private}} = \mathbb{E}_{\boldsymbol{s}\in\mathcal{D}, c\in C, \boldsymbol{u}^{(c)}\in\boldsymbol{\phi}(\boldsymbol{s})}\ \mathrm{InfoNCE}(\boldsymbol{u}^{(c)}, \boldsymbol{u}_+^{(c)}, \boldsymbol{u}_-^{(c)}). \tag{7}$$

Both loss functions can be combined into the disentanglement loss

$$\mathcal{L}_{\text{disent}} = \mathcal{L}_{\text{shared}} + \mathcal{L}_{\text{private}}, \tag{8}$$

which ensure maximization of similarity among the shared representation, minimization of similarity between shared and private representations and minimization of similarity between individual private representations. Apart from the contrastive objective, DisDP adds an orthogonality loss

$$\mathcal{L}_{\text{ortho}} = \mathbb{E}_{\boldsymbol{s}\in\mathcal{D}, c\in C, (\boldsymbol{v}^{(c)}, \boldsymbol{u}^{(c)}), \in\boldsymbol{\phi}(\boldsymbol{s})} \langle \boldsymbol{v}^{(c)}, \boldsymbol{u}^{(c)} \rangle^2, \tag{9}$$

to further disentangle the shared and private embeddings by minimizing the squared dot product $\langle\cdot,\cdot\rangle$ between them for each camera. Together with the diffusion loss, this results in the final loss

$$\mathcal{L} = \mathcal{L}_{\text{diffusion}} + \lambda_{\text{disent}} \cdot \mathcal{L}_{\text{distent}} + \lambda_{\text{ortho}} \cdot \mathcal{L}_{\text{ortho}}, \tag{10}$$

where λ_{disent} and λ_{ortho} are hyperparameters scaling the importance of the disentanglement and orthogonality loss.

IV. Evaluation

The experiments conducted in this paper try to answer 3 research questions increasing in difficulty towards robustness and 1 research question with focus on interpretability:

RQ1: Does disentanglement affect the performance of IL policies?

RQ2: Do disentangled latent spaces improve resilience to noisy sensor input and complete sensor dropout?

RQ3: How resilient are policies to environmental perturbations and sensor dropout?

RQ4: Does disentanglement results in more interpretable latent spaces?

To answer these questions all policies are evaluated on two sota IL benchmarks, The Colosseum [31] and Libero [22]. Both environments provide multi-camera image observations with 5 cameras for The Colosseum and 2 cameras for Libero. The Colosseum is constructed using tasks from RLBench [12], to benchmark complex robot manipulation tasks. It has 20 tabletop tasks with different variations in each task, including changes in lighting, texture, object colors and properties. Libero consists of diverse robot manipulation tasks categorized into object, spatial, goal, and long-horizon tasks. These tasks evaluate robotic skills on different skill ranges, making it a comprehensive benchmark for generalization in robotic manipulation. In both benchmarks, Policy performance is assessed using success rate, defined as the percentage of rollouts that successfully complete the task within a specified number of steps.

A. Evaluated approaches

During the evaluation, three baselines are considered:

BC: Behavior Cloning (BC) is usually used as a default baseline for imitation learning. We apply an encoder-decoder Transformer architecture to perform action prediction, which is optimized by Mean Squared Error (MSE).

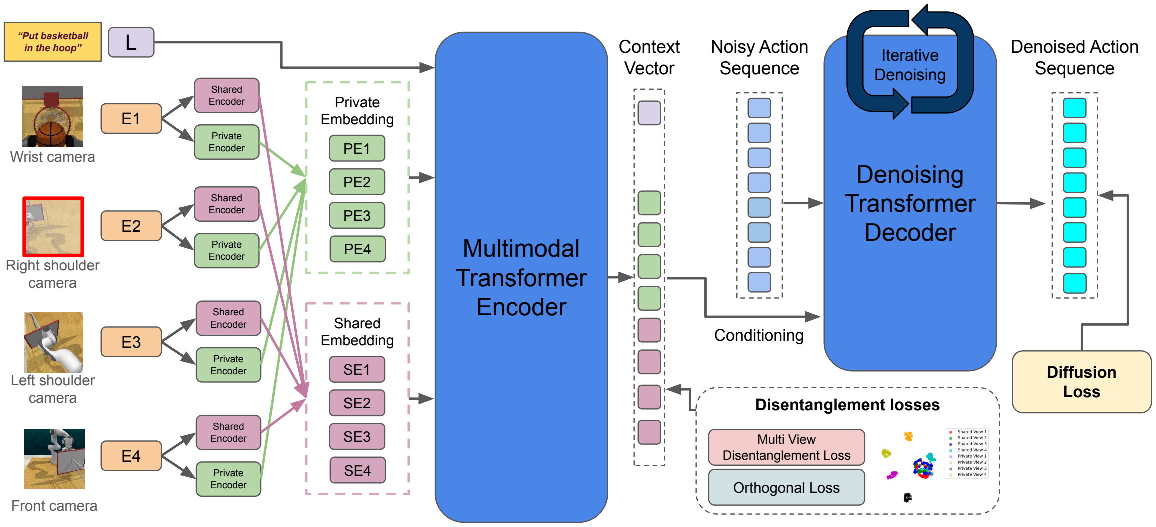

Fig. 2: Overview of Disentangled Diffusion Policy (DisDP). The model processes multi-view image inputs by separating them into shared and private representations. The language instruction is encoded using Clip and each camera input is encoded using ResNet-18, followed by disentanglement modules that extract shared embeddings across all views and private embeddings for individual views. These embeddings are processed by a multimodal transformer encoder and serve as conditioning inputs to the denoising transformer decoder for action prediction. The model is trained with a combination of diffusion loss, multi-view disentanglement loss, and orthogonality loss to enforce representation separation. This structured representation learning enhances robustness to sensor noise, failures, and environmental variations.

BESO-ACT: BEhavior generation with ScOre-based Diffusion Policies (BESO) [32] is a diffusion-based policy that represents the denoising process using a continuous Stochastic-Differential Equation (SDE). Beyond that, we build BESO-ACT by using the same Transformer in BC and applying action chunking [46].

BESO-ACT-dropout: This baseline uses BESO-ACT but introduces random modality dropout in training at a rate of 10 percent to gain robustness.

Our Contributed methods are:

DisBC: DisBC extends the BC baseline by introducing disentangled latent spaces.

DisDP: DisDP integrates disentangled representations in the BESO-ACT architecture

B. Experimental Setup

The Colosseum: With regards to **RQ1**, experiments are conducted on 10 of the 20 Colosseum tasks: basketball in hoop, close box, close laptop lid, hockey, meat on grill, move hanger, open drawer, reach and drag, scoop with spatula, and slide block to target. These tasks were selected based on their strong performance using the baseline method, ensuring a fair comparison. The proposed methods and baselines are trained on the *no-variation* setting within the Colosseum suite for 200 epochs on the same hyperparameters to avoid biases. The trained models are evaluated on noisy camera sensor input and complete dropout to address **RQ2**. Regarding **RQ3**, the trained models are evaluated on 8 different Colosseum variations: no-variation, background texture, camera pose, distractor, light color, object color, table color, table texture. The dataset contains 100 demonstrations for each task with images captured from the five camera views. The policies are evaluated using three seeds, with 25 rollouts per task and a maximum of 300 steps per rollout.

Libero: Addressing **RQ1** policies are evaluated on 3 of the 4 categories, excluding long-horizon tasks for computation reasons. Models are trained for 50 epochs on 60 percent of demonstrations on same hyperparameters to ensure fair comparison. Libero includes 2 camera views: Agent camera **0** and in-hand camera **1**. Regarding **RQ2**, policies are evaluated on dropping out either the agent or in-hand view. The methods are evaluated using three different seeds with 25 rollouts per task in each dataset split. Each episode has maximum 260 steps per rollout.

C. Result analysis

The following section discusses the 4 introduced research questions with regard to the experimental results on The Colosseum [31] and Libero [22] benchmarks.

RQ1: Does disentanglement affect the performance of IL policies?

The first research question aims at analyzing the quality of policies when adding disentanglement, because of the trade-off between performance and interpretability. The row *None* in Table I displays the results for all three baselines and the two proposed methods using disentangled shared and private embeddings. In both benchmarks, using the disentangled version of the baseline does improve overall performance. DisDP achieves 0.896 success rate compared to the 0.709 of BESO-

ACT on the Colosseum tasks. It also improves results on the Libero benchmark between 0.06 and 0.12, compared to the BESO-ACT baseline. In general, disentangled IL policies do improve overall performance.

View(s)		**BC**	**DisBC**	**BESO-ACT**	**BESO-ACT-Dropout**	**DisDP**
None		0.361 ± 0.11	0.540 ± 0.08	0.709 ± 0.03	0.435 ± 0.04	**0.896** ± **0.05**
0	Noisy	0.160 ± 0.05	0.444 ± 0.04	0.000 ± 0.00	0.020 ± 0.02	**0.568** ± **0.11**
	Masked	0.096 ± 0.01	0.206 ± 0.03	0.068 ± 0.05	0.096 ± 0.01	**0.440** ± **0.03**
1	Noisy	0.028 ± 0.03	0.496 ± 0.05	0.288 ± 0.07	0.326 ± 0.07	**0.500** ± **0.12**
	Masked	0.120 ± 0.02	0.140 ± 0.03	0.196 ± 0.04	0.168 ± 0.02	**0.632** ± **0.04**
2	Noisy	0.100 ± 0.02	0.196 ± 0.03	0.008 ± 0.01	0.280 ± 0.01	**0.306** ± **0.08**
	Masked	0.048 ± 0.01	0.228 ± 0.01	0.292 ± 0.03	0.100 ± 0.03	**0.420** ± **0.02**
3	Noisy	0.130 ± 0.02	**0.440** ± **0.02**	0.252 ± 0.03	0.210 ± 0.07	0.280 ± 0.04
	Masked	0.028 ± 0.02	**0.096** ± **0.01**	0.040 ± 0.03	0.004 ± 0.00	0.060 ± 0.03
0 1	Noisy	0.020 ± 0.01	**0.420** ± **0.01**	0.000 ± 0.00	0.020 ± 0.01	0.378 ± 0.05
	Masked	0.056 ± 0.01	0.100 ± 0.02	0.028 ± 0.02	0.048 ± 0.01	**0.196** ± **0.05**
1 2	Noisy	0.080 ± 0.07	**0.370** ± **0.02**	0.000 ± 0.00	0.186 ± 0.04	0.172 ± 0.04
	Masked	0.000 ± 0.00	0.092 ± 0.01	0.070 ± 0.01	0.040 ± 0.02	**0.192** ± **0.07**

TABLE I: **Colosseum no variation Dataset Evaluation with Noisy and Masked Camera Views**. The numbers in the column *View(s)* correspond to the specific camera: **0** left view, **1** right view, **2** wrist view, and **3** front view. Dual camera dropouts are only reported for **0 1** and **1 2** because other combinations achieve low success rate for all methods. The evaluation examines how noisy sensors and sensor failures affect task success rates and assesses the resilience of different methods under these conditions. The disentangled methods perform much better compared to their baseline implementations. Especially the DisBC has a small decrease in performance, when adding noise.

RQ2: Do disentangled latent spaces improve resilience to noisy sensor input and complete sensor dropout?

For all methods, the overall performance on The Colosseum benchmark does drop significantly, as shown in the *Noisy* rows of Table I. The disentangled methods still outperform their corresponding baseline methods and with less performance loss. Especially the DisBC still performs similar to the non-noisy results. The traditional BC, on the other hand, completely fails when confronted with noisy sensor inputs. In the noisy scenario, the BESO-ACT-dropout is also able to retain more of its original performance, compared to the BESO-ACT.

As shown in Table II and the *Masked* rows in Table I, across both Libero and Colosseum, BESO-ACT and BESO-ACT-Dropout experience significant performance drops when critical camera views are unavailable, highlighting their reliance on complete visual input. Notably, BESO-ACT-Dropout fails to mitigate sensor failures, showing that naive modality dropout during training does not improve robustness but instead leads to the loss of important task-relevant information. In Colosseum, evaluation is conducted with four cameras, providing redundancy and robustness to sensor failures due to overlapping viewpoints. In contrast, Libero only has two cameras in total, relying heavily on both of them. Unlike BC and BESO-ACT, the disentangled-based methods DisBC and DisDP exhibit greater resilience to sensor failures, maintaining higher success rates across masked conditions. DisDP consistently outperforms all baselines, even when multiple modalities are missing.

Masked		**BC**	**DisBC**	**BESO-ACT**	**BESO-ACT-Dropout**	**DisDP**
None	Object	0.684 ± 0.00	0.736 ± 0.02	0.752 ± 0.00	0.514 ± 0.05	**0.816** ± **0.02**
	Spatial	0.556 ± 0.00	0.583 ± 0.02	0.580 ± 0.03	0.552 ± 0.04	**0.701** ± **0.04**
	Goal	-	-	0.576 ± 0.02	0.418 ± 0.05	**0.680** ± **0.09**
0	Object	0.000 ± 0.00	0.110 ± 0.03	0.204 ± 0.00	0.004 ± 0.00	**0.295** ± **0.04**
	Spatial	0.000 ± 0.00	0.000 ± 0.00	0.028 ± 0.00	0.023 ± 0.00	**0.144** ± **0.02**
	Goal	-	-	**0.084** ± **0.01**	0.040 ± 0.00	0.004 ± 0.00
1	Object	0.000 ± 0.00	0.000 ± 0.00	0.012 ± 0.01	0.000 ± 0.00	**0.226** ± **0.03**
	Spatial	0.000 ± 0.00	0.004 ± 0.00	0.004 ± 0.00	0.023 ± 0.04	**0.112** ± **0.00**
	Goal	-	-	0.012 ± 0.00	0.004 ± 0.00	**0.200** ± **0.04**

TABLE II: **Libero dataset evaluation**: The evaluation examines three task suites—Object, Spatial, and Goal—across three conditions: normal (all cameras available), agent view camera masked (0), and in-hand camera masked (1). The results demonstrate the effect of modality dropout on task success and highlight that policies trained with disentangled methods exhibit better adaptability to missing sensory inputs.

DisDP achieves the highest performance retention under modality dropout, demonstrating that disentangled representations effectively preserve task-relevant features despite missing inputs. While DisBC also leverages shared and private representation separation and improves robustness over BC, it does not match the adaptability of DisDP, which benefits from diffusion policies in addition to disentangled representations. This analysis shows that disentangled IL policies are much more resilient towards noisy sensor inputs and that disentangled representations help retain important task information and handle sensor dropout.

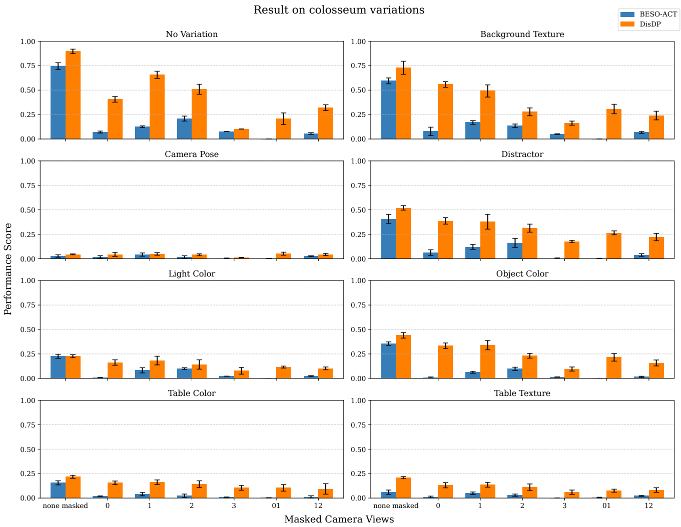

Fig. 3: **Colosseum results on variations, comparison between BESO-ACT and DisDP**. The no-variation condition serves as a baseline, showing the highest performance. Spatial, textural, and lighting variations significantly impact success rates, with camera pose and table texture masking causing the most degradation.

RQ3: How resilient are policies to environmental perturbations and sensor dropout?

To evaluate the robustness of policies on environmental perturbations and modality dropouts, Colosseum provides 7 different variations for all tasks. Previous experiments showed that the diffusion-based methods perform best on The Colosseum, therefore only those two methods are evaluated on the variations of The Colosseum.

Figure 3 presents the evaluations on the environmental variations from The Colosseum. The no-variation condition serves as the baseline, achieving the highest performance across all tasks. The results indicate the degradation in performance on environmental perturbations. The performance decreases further when certain camera views are dropped out. Overall, DisDP demonstrates greater robustness compared to BESO-ACT, particularly in handling object color, table color, and background texture variations.

The results show that disentangled representations help handle environmental changes. Even without training on these variations, our method remained more robust. Variations in camera pose, table texture, and lighting were the most disruptive factors, affecting the model's spatial reasoning and fine-grained perception. These observations address **RQ3**, as it highlights how disentangled representations enable improved generalization to environmental perturbations by preserving task-relevant features while filtering out irrelevant variations.

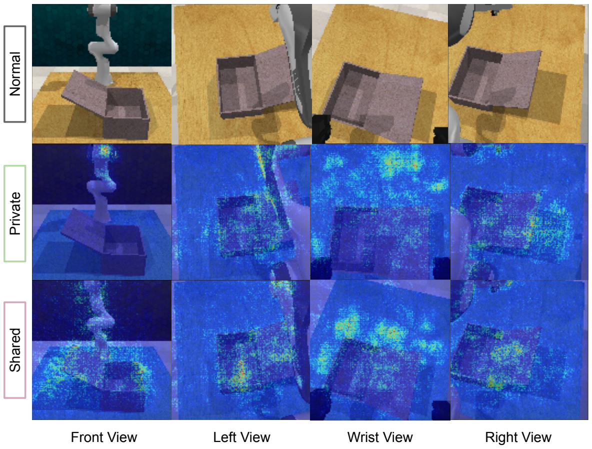

Fig. 4: **Saliency maps for disentangled embeddings**. In the close box task, the shared embeddings capture the box edges, which are crucial for task completion and visible across different views. In contrast, the private embeddings focus on specific details, such as robot joints and table shadows, which contribute to task execution, while others capture unique but less relevant scene elements.

RQ4: Does disentanglement results in more interpretable latent spaces?

DisDP's superior performance in Libero and Colosseum shows that separating shared and private components enhances robustness and adaptability under both normal and unreliable camera conditions. To investigate the interpretability of the disentangled latent space, we examine the saliency maps of the learned shared and private representations in Figure 4, using the close-box task as an example. The shared representation focuses on box edges, a crucial feature for proper alignment and closure, ensuring that essential task information remains consistent across views. This cross-view consistency allows the model to retain key information, even when some camera inputs are missing or degraded. In contrast, the private representations capture view-specific details, such as robot joints and table shadows, which provide additional contextual information for precise manipulation.

V. Conclusion

This work introduced Disentangled Diffusion Policy (DisDP), a method for improving robustness in IL by leveraging multi-view disentanglement. By structuring sensor inputs into shared and private representations, DisDP enhances the model's ability to handle sensor noise, dropouts and environmental variations better. Our evaluations on The Colosseum and Libero demonstrate that disentangled methods achieve better performance than their baseline implementation, even when all sensors are available.

Evaluations with noisy or unreliable sensors demonstrated the robustness improvement through disentangled IL methods. Furthermore, disentanglement additionally provides more robustness towards environmental changes, making models more robust in general. The separation of private and shared embeddings allows for visualization of the latent space through Gradient-weighted Class Activation Mapping (Grad-CAM) and Uniform Manifold Approximation and Projection (UMAP). These visualizations give insight into the focus of the model and how private and shared embeddings are separated.
Limitations of DisDP can be observed when looking at camera dropout combinations. If specific combinations of cameras are not available, disentanglement does not retain information to complete tasks reliably. Furthermore, less cameras decrease the efficiency of disentangled methods, because the shared embedding has less overlap between viewpoints.
Future work will focus on improving robustness under less modality inputs and to reduce performance loss if more then one modality is not available. The next steps will also include real robot experiments, to confirm the proposed method outside of simulation. Furthermore, including other sensor modalities beside vision would be interesting and could enhance model performance.

Acknowledgments

The research presented in this paper was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–448648559. The authors also acknowledge support by the state of Baden-Wurttemberg through the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Wurttemberg and by the German Federal Ministry of Education and Research.

References

[1] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from

demonstration. *Robotics and autonomous systems*, 57 (5):469–483, 2009.

[2] Philipp Becker, Sebastian Mossburger, Fabian Otto, and Gerhard Neumann. Combining reconstruction and contrastive methods for multimodal representations in rl. In *Reinforcement Learning Conference*, 2024.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[4] 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. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[5] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *International conference on machine learning*, pages 1282–1289, California, 2019. PMLR.

[6] Atalay Donat, Xiaogang Jia, Xi Huang, Aleksandar Taranovic, Denis Blessing, Ge Li, Hongyi Zhou, Hanyi Zhang, Rudolf Lioutikov, and Gerhard Neumann. Towards fusing point cloud and visual representations for imitation learning. *arXiv preprint arXiv:2502.12320*, 2025.

[7] Mhairi Dunion and Stefano V Albrecht. Multi-view disentanglement for reinforcement learning with multiple cameras. *arXiv preprint arXiv:2404.14064*, 2024.

[8] Yilun Hao, Ruinan Wang, Zhangjie Cao, Zihan Wang, Yuchen Cui, and Dorsa Sadigh. Masked imitation learning: Discovering environment-invariant modalities in multimodal demonstrations. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–7, 2023. doi: 10.1109/IROS55552.2023.10341728.

[9] Yilun Hao, Ruinan Wang, Zhangjie Cao, Zihan Wang, Yuchen Cui, and Dorsa Sadigh. Masked imitation learning: Discovering environment-invariant modalities in multimodal demonstrations. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–7. IEEE, 2023.

[10] Ryan Hoque, Ajay Mandlekar, Caelan Garrett, Ken Goldberg, and Dieter Fox. Intervengen: Interventional data generation for robust and data-efficient robot imitation learning. *arXiv preprint arXiv:2405.01472*, 2024.

[11] Nihal Jain, Praneetha Vaddamanu, Paridhi Maheshwari, Vishwa Vinay, and Kuldeep Kulkarni. Self-supervised multi-view disentanglement for expansion of visual collections. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 841–849, 2023.

[12] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

[13] Xiaogang Jia, Denis Blessing, Xinkai Jiang, Moritz Reuss, Atalay Donat, Rudolf Lioutikov, and Gerhard Neumann. Towards diverse behaviors: A benchmark for imitation learning with human demonstrations. *arXiv preprint arXiv:2402.14606*, 2024.

[14] Xiaogang Jia, Atalay Donat, Xi Huang, Xuan Zhao, Denis Blessing, Hongyi Zhou, Hanyi Zhang, Han A Wang, Qian Wang, Rudolf Lioutikov, et al. X-il: Exploring the design space of imitation learning policies. *arXiv preprint arXiv:2502.12330*, 2025.

[15] Yiding Jiang, J. Zico Kolter, and Roberta Raileanu. On the importance of exploration for generalization in reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 12951–12986, New Orleans, USA, 2023. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2a4310c4fd24bd336aa2f64f93cb5d39-Paper-Conference.pdf.

[16] Joshua Jones, Oier Mees, Carmelo Sferrazza, Kyle Stachowicz, Pieter Abbeel, and Sergey Levine. Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding. *arXiv preprint arXiv:2501.04693*, 2025.

[17] Guanzhou Ke, Yang Yu, Guoqing Chao, Xiaoli Wang, Chenyang Xu, and Shengfeng He. Disentangling multi-view representations beyond inductive bias. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2582–2590, 2023.

[18] Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76:201–264, 2023.

[19] Michelle A Lee, Matthew Tan, Yuke Zhu, and Jeannette Bohg. Detect, reject, correct: Crossmodal compensation of corrupted sensors. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 909–916. IEEE, 2021.

[20] Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A Lee, Huazhe Xu, Edward Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. *arXiv preprint arXiv:2212.03858*, 2022.

[21] Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36:32971–32998, 2023.

[22] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] Rui Liu, Amisha Bhaskar, and Pratap Tokekar. Adaptive visual imitation learning for robotic assisted feeding

across varied bowl configurations and food types. *arXiv preprint arXiv:2403.12891*, 2024.

[24] Zeyi Liu, Cheng Chi, Eric Cousineau, Naveen Kuppuswamy, Benjamin Burchfiel, and Shuran Song. Maniwav: Learning robot manipulation from in-the-wild audio-visual data. In *8th Annual Conference on Robot Learning*, 2024.

[25] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310, 2021.

[26] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *5th Annual Conference on Robot Learning*, 2021. URL https://openreview.net/forum?id=JrsfBJtDFdI.

[27] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2021.

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[29] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2):1–179, 2018.

[30] Jyothish Pari, Nur Muhammad Shafiullah, Sridhar Pandian Arunachalam, and Lerrel Pinto. The surprising effectiveness of representation learning for visual imitation. *arXiv preprint arXiv:2112.01511*, 2021.

[31] Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation. *arXiv preprint arXiv:2402.08191*, 2024.

[32] Moritz Reuss, Maximilian Li, Xiaogang Jia, and Rudolf Lioutikov. Goal-conditioned imitation learning using score-based diffusion policies. In *Robotics: Science and Systems*, 2023.

[33] Moritz Reuss, Jyothish Pari, Pulkit Agrawal, and Rudolf Lioutikov. Efficient diffusion transformer policies with mixture of expert denoisers for multitask learning. *arXiv preprint arXiv:2412.12953*, 2024.

[34] Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. In *Robotics: Science and Systems*, 2024.

[35] Nur Muhammad Shafiullah, Zichen Cui, Ariuntuya Arty Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone. *Advances in neural information processing systems*, 35:22955–22968, 2022.

[36] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.

[37] Skand Skand, Bikram Pandit, Chanho Kim, Li Fuxin, and Stefan Lee. Simple masked training strategies yield control policies that are robust to sensor failure. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=AsbyZRdqPv.

[38] Abitha Thankaraj and Lerrel Pinto. That sounds right: Auditory self-supervision for dynamic robot manipulation. In *Conference on Robot Learning*, pages 1036–1049. PMLR, 2023.

[39] Miguel Vasco, Hang Yin, Francisco S Melo, and Ana Paiva. How to sense the world: Leveraging hierarchy in multimodal perception for robust reinforcement learning agents. *arXiv preprint arXiv:2110.03608*, 2021.

[40] Miguel Vasco, Hang Yin, Francisco S Melo, and Ana Paiva. Leveraging hierarchy in multimodal generative models for effective cross-modality inference. *Neural Networks*, 146:238–255, 2022.

[41] Annie Xie, Lisa Lee, Ted Xiao, and Chelsea Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3153–3160. IEEE, 2024.

[42] TengQi Ye, Tianchun Wang, Kevin McGuinness, Yu Guo, and Cathal Gurrin. Learning multiple views with orthogonal denoising autoencoders. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part I 22*, pages 313–324. Springer, 2016.

[43] Zhecheng Yuan, Tianming Wei, Shuiqi Cheng, Gu Zhang, Yuanpei Chen, and Huazhe Xu. Learning to manipulate anywhere: A visual generalizable framework for reinforcement learning. *arXiv preprint arXiv:2407.15815*, 2024.

[44] Maryam Zare, Parham M. Kebria, Abbas Khosravi, and Saeid Nahavandi. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*, 54(12):7173–7186, 2024. doi: 10.1109/TCYB.2024.3395626.

[45] Amy Zhang, Nicolas Ballas, and Joelle Pineau. A dissection of overfitting and generalization in continuous reinforcement learning. *arXiv preprint arXiv:1806.07937*, 1 (1), 2018.

[46] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.