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 large-scale deployment, robots must be robust to perturbations such as environmental variations, sensor noise, and unavailability of sensors at test time that were available during training (sens. While environmental robustness has been explored [27], sensor dropout remains understudied. Existing methods often fail on complex, multi-view benchmarks [32, 9], exposing a critical vulnerability in current IL-based policies. To address this, we propose **Disentangled Diffusion Policy (DisDP)**, a method that disentangles sensor modalities into shared and private embeddings. While integrating multiple sensors improves robustness to noisy or unreliable inputs [17, 31, 33, 21, 13], such setups are prone to calibration errors, occlusion, and failures. Most approaches assume fully reliable inputs [33, 29], limiting robustness to dropout.

This work focuses on vision-based IL and resilience to missing or noisy camera inputs (Figure 1). IL is widely used for acquiring complex behaviors [1, 25], with recent multi-task methods showing strong performance [26, 30, 31, 9, 29, 28, 4, 6]. Despite the progress, most IL methods rely on latent spaces not designed for degraded input, making them vulnerable to sensor failures. We propose a disentangled representation approach for IL, separating inputs into shared and private embeddings to improve robustness and interoperability. We apply this to both a diffusion policy [28, 29] and a Transformer-based BC model [23], showing improved performance and reduced degradation under noisy and incomplete sensing.

II. Related Work

a) Robustness in Behavior Learning: Behavior learning suffers from generalization issues, often resulting in performance degradation in unfamiliar environments due to overfitting and limited adaptability [41, 5, 12, 15, 40]. To address this, various methods have been proposed to improve robustness under modality dropout [26, 39, 20, 37, 10, 2].One line of work focuses on estimating missing modalities. For example, SMIL [22] uses Bayesian meta-learning with variational inference, while CCM [16] applies self-supervision to filter and reconstruct corrupted inputs. However, these methods do not address complete sensory failure.

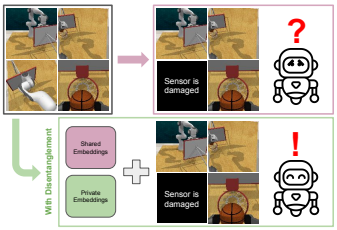

Fig. 1: Robotic policies rely on multiple sensory inputs, making them vulnerable to failures. This work explores how disentangling sensory data into shared and private embeddings improves robustness under sensor dropouts.

Several approaches handle missing or irrelevant modalities to enhance robustness. Masking methods suppress irrelevant modalities [32, 9], while MIL [8] masks before policy creation without addressing sensor failures. Hierarchical methods like Nexus [35] average embeddings, limiting expressiveness, whereas MUSE [34] uses Product-of-Experts for better integration. For multi-view cameras, RL-based disentanglement improves robustness [7]. In contrast, DisDP uses contrastive disentanglement for imitation learning, tested on Colosseum and Libero.

b) Multi View Disentanglement: Multi-view disentanglement separates features into shared and private components across modalities. Orthogonal [38], self-supervised [11, 14], and information-theoretic methods [18] achieve this via orthogonality, contrastive losses, or mutual information. These approaches improve generalization by isolating task-relevant features and suppressing noise.

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

This work addresses 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. The goal is to develop a framework that robustly handles 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 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 IL approach. In a multi-modal IL setting the state information contains multiple modalities, typically across different sensors. This work includes:

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

To model sensor noise and availability, we define **reliability masks** $\boldsymbol{M}_k = (M_k^{(1)}, M_k^{(2)}, \ldots, M_k^{(C)})$. $M_k = \mathbf{1}$ 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. While individual embeddings can theoretically improve robustness to modality dropout, learned policies often still rely on all modalities being present and reliable. In contrast, this work explicitly learns shared embeddings $\boldsymbol{v}$ across sensors and private embeddings $\boldsymbol{u}$ per sensor, rather than individual or unified embeddings.

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

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.

Recent work on action chunking [42] shows that predicting action sequences outperforms single-step prediction. Following this, 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

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

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.

B. Disentangled Diffusion Policy

Disentangled Diffusion Policy (DisDP) combines a Transformer based encoder-decoder diffusion model [28, 29] 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 embeddings capture global, view-consistent features, providing robustness when cameras are occluded or unreliable. The private embeddings retain fine-grained, view-specific details that enrich the policy when available.

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 [24, 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

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

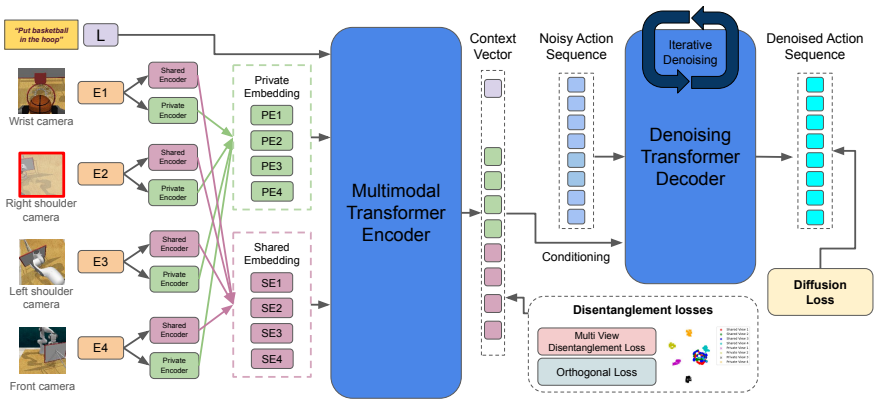

Fig. 2: **Overview of Disentangled Diffusion Policy (DisDP).** The model processes multi-view image inputs by separating them into shared and private representations. Language instructions are encoded with CLIP, and each camera view with ResNet-18, followed by disentanglement modules that extract shared embeddings across views and private ones per view. These embeddings are processed by a multimodal transformer encoder and used to condition a denoising transformer decoder for action prediction. The model is trained using diffusion, multi-view disentanglement, and orthogonality losses to enforce representation separation. This structured representation enhances robustness to sensor noise, failures, and environmental changes.

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 [36], 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

This work evaluates four research questions related to robustness and interpretability, using two state-of-the-art IL benchmarks: The Colosseum [27] and Libero [19]. Policy performance is measured as success rate—the percentage of rollouts that complete the task within a fixed number of steps.

A. Evaluated Approaches

The following three baselines are evaluated: **BC**, a standard BC baseline using a Transformer-based encoder-decoder for action prediction, trained with MSE; **BESO-ACT**, a diffusion policy based on BEhavior generation with ScOre-based Diffusion Policies (BESO) [28], using a continuous Stochastic-Differential Equation (SDE) and the same Transformer as BC, combined with action chunking [42]; and **BESO-ACT-dropout**, improving robustness of BESO-ACT by applying random modality dropout (10%) during training.

Our contributed methods are **DisBC**, extending BC with disentangled latent spaces, and **DisDP**, integrating disentangled representations in the BESO-ACT architecture.

B. Experimental Setup

The Colosseum: Experiments are conducted on 10 tasks selected based on strong baseline performance to ensure fair and meaningful comparison. All methods are trained for 200 epochs in the *no-variation* setting using identical hyperparameters. Evaluation is performed under camera noise and sensor dropout conditions, as well as across 8 visual variations: no-variation, background texture, camera pose, distractor, light color, object color, table color, and table texture. Each task includes 100 demonstrations with images from five camera views. Policies are evaluated using three random seeds, with 25 rollouts per task and a maximum of 300 steps per rollout.

Libero: Policies are evaluated on three categories, excluding long-horizon tasks due to computational constraints. All models are trained for 50 epochs using 60% of demonstrations and identical hyperparameters. Libero provides two camera views: an agent view (camera 0) and an in-hand view (camera 1). Evaluation includes single-view dropout scenarios. Each method is tested with three random seeds, 25 rollouts per task, and a maximum of 260 steps per rollout.

C. Result analysis

RQ1: Does disentanglement affect the performance of IL policies? As shown in the *None* rows of Table I and II, using disentangled shared and private embeddings improves performance across both benchmarks. **RQ2: Do disentangled latent spaces improve resilience to noisy sensor input and complete sensor dropout?** As shown in the *Noisy*

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 dataset evaluation**. 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.

rows of Table I, all methods experience performance drops on Colosseum when adding noise to the camera inputs, but disentangled methods degrade less and consistently outperform their baselines. Across both Libero and Colosseum (Table II, *Masked* rows in Table I), 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.

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).

RQ3: How resilient are policies to environmental perturbations and sensor dropout? The results (Figure 3) indicate the degradation in performance on environmental perturbations for all methods. 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.

RQ4: Does disentanglement results in more interpretable latent spaces? 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. In the close-box task, the shared embeddings capture the box edges, which are

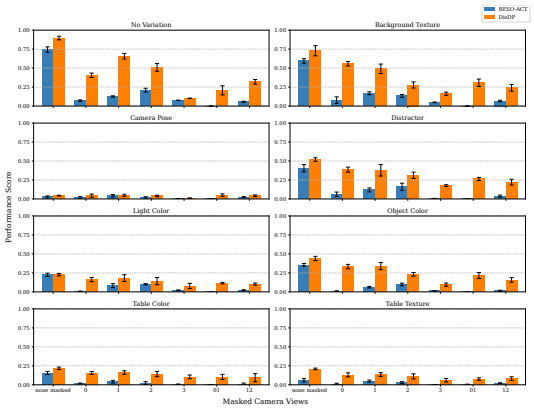

Fig. 3: **Colosseum results on variations**. Previous experiments showed that the diffusion-based methods perform best on The Colosseum, therefore only those two methods are evaluated on the variations of The Colosseum.

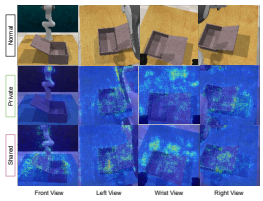

Fig. 4: **Exemplary saliency maps for disentangled embeddings.**

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.

V. Conclusion

This work introduced Disentangled Diffusion Policy (DisDP), a method for improving robustness in IL through multi-view disentanglement. By structuring sensor inputs into shared and private representations, DisDP improves resilience to sensor noise, dropouts, and environmental variations. Evaluations on The Colosseum and Libero benchmarks show that disentangled methods outperform their baselines, even when all sensors are available. Further experiments under noisy and missing input conditions confirm that disentanglement enhances robustness across sensor and environmental shifts. The separation of shared and private embeddings also enables Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations, offering insights into model focus and latent structure.

While DisDP improves robustness overall, performance degrades when key camera combinations are missing, especially with fewer available views reducing the effectiveness of shared embeddings. Future work will focus on improving performance with limited modalities, validating the approach on real-world robots, and extending it to additional sensor types beyond vision.

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] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7): 1661–1674, 2011. doi: 10.1162/NECO_a_00142.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.