Shortcut Learning in Generalist Robot Policies: The Role of Dataset Diversity and Fragmentation

Youguang Xing[1*], **Xu Luo**[1*], **Junlin Xie**[1], **Lianli Gao**[1], **Hengtao Shen**[2], **Jingkuan Song**[2†]
[1]UESTC, [2]Tongji University
ygxing@std.uestc.edu.cn, frank.luox@outlook.com

Abstract: Generalist robot policies trained on large-scale datasets such as Open X-Embodiment (OXE) demonstrate strong performance across a wide range of tasks. However, they often struggle to generalize beyond the distribution of their training data. In this paper, we investigate the underlying cause of this limited generalization capability. We identify *shortcut learning*—the reliance on task-irrelevant features—as a key impediment to generalization. Through comprehensive theoretical and empirical analysis, we uncover two primary contributors to shortcut learning: (1) limited diversity within individual sub-datasets, and (2) significant distributional disparities across sub-datasets, leading to dataset fragmentation. These issues arise from the inherent structure of large-scale datasets like OXE, which are typically composed of multiple sub-datasets collected independently across varied environments and embodiments. Our findings provide critical insights into dataset collection strategies that can reduce shortcut learning and enhance the generalization ability of generalist robot policies. Moreover, in scenarios where acquiring new large-scale data is impractical, we demonstrate that carefully selected robotic data augmentation strategies can effectively reduce shortcut learning in existing offline datasets, thereby improving generalization capabilities of generalist robot policies, e.g., π_0, in both simulation and real-world environments. More information at our website[1].

Keywords: Generalist Robot Policies, Shortcut Learning, Large-Scale Robot Datasets

1 Introduction

The recent advancements in machine learning, particularly in domains such as computer vision and natural language processing, can be largely attributed to the scaling up of both data and model sizes. Notably, scaling laws in these domains [1, 2, 3] indicate a consistent trend of performance improvement and emergent generalization capabilities as the number of model parameters and the volume of data are increased.

It is anticipated that analogous trends will emerge in the field of robotics. Consequently, recent research efforts in the field of robot learning have concentrated on the development of increasingly large-scale robot datasets [6, 5, 8, 9, 10, 11] and the training of high-capacity models [6, 12, 13, 14, 15, 7, 16, 17] on these datasets, which directly map observations to actions, e.g., Vision-Language-Action (VLA) models [12]. The hope is that, by feeding abundant web and robot data, we can develop a generalist robot policy capable of addressing a wide spectrum of tasks and, more importantly, generalizing to novel tasks and environments out of box.

Despite advancements in training models on large-scale datasets like Open X-Embodiment (OXE) [10], these models continue to demonstrate limited generalization capabilities across multiple axes,

[*]: denotes equal contribution. [†]: denotes corresponding author.

[1] Poject page: https://lucky-light-sun.github.io/proj/shortcut-learning-in-grps/

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

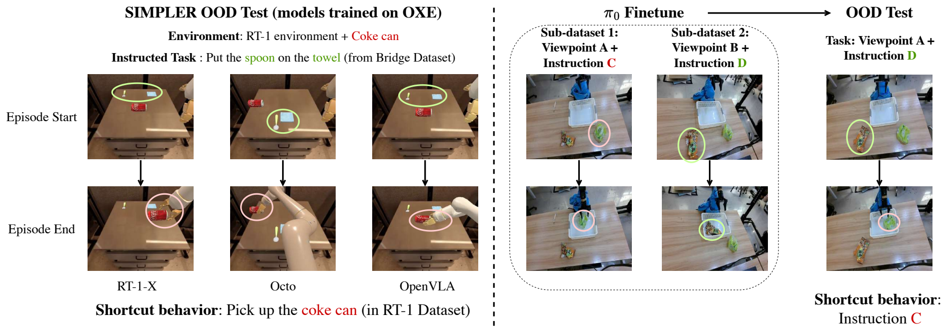

SIMPLER OOD Test (models trained on OXE)
Environment: RT-1 environment + Coke can
Instructed Task : Put the spoon on the towel (from Bridge Dataset)
Episode Start
Episode End
RT-1-X
Octo
OpenVLA
Shortcut behavior: Pick up the coke can (in RT-1 Dataset)
π_0 Finetune
OOD Test
Sub-dataset 1: Viewpoint A + Instruction C
Sub-dataset 2: Viewpoint B + Instruction D
Task: Viewpoint A + Instruction D
Shortcut behavior: Instruction C

Figure 1: **Demonstrations of shortcut learning in generalist robot policies. Left:** Three generalist robot policies trained on the OXE dataset exhibit shortcut behavior in the SIMPLER environment [4]. Despite being tasked with "put the spoon on the towel", a task present in the Bridge sub-dataset [5], all models consistently perform the task "pick up the coke" which is exclusive to the RT-1 sub-dataset [6]. **Right:** π_0 [7] policy after finetuning on real-world data exhibits shortcut behavior. The policy was finetuned on two distinct data subsets: (Viewpoint A, Instruction C) and (Viewpoint B, Instruction D). When tasked with instruction D from the novel configuration of Viewpoint A, the policy incorrectly executes Instruction C. This indicates that the policy has learned to associate the viewpoint with the action, ignoring the provided instruction.

including visual, semantic, and behavioral aspects [18]. This limitation cannot be ascribed to a deficiency in data, as the scale of OXE—comprising over one million episodes—surpasses that of datasets typically employed for training vision-language models, which generally consist of fewer than one million images and yet exhibit strong generalization capabilities [19]. *So, what hinders generalization in robot policies?*

In this paper, we identify *shortcut learning*—a model's reliance on spurious correlations between actions and task-irrelevant components of observations—as a significant contributor to this limitation in generalization. As illustrated in Figure 1, by learning from confounding factors such as viewpoint, background, and texture, the model fails to capture the true causal relationships between observations and actions. Consequently, it may overlook essential elements like language instructions and target objects, thereby restricting its ability to generalize beyond the training distribution.

To investigate the root causes of shortcut learning in generalist robot policies, we conduct a detailed analysis of the widely used OXE dataset. Our visual and textual feature analysis reveals two critical issues: (1) limited diversity within sub-datasets, and (2) significant disparities between them, resulting in dataset *fragmentation*. Through theoretical analysis and controlled experiments, we demonstrate that both characteristics contribute to shortcut learning. Based on these findings, we derive key insights for improved robot dataset collection strategies, summarized below:

1. Ensure diversity in both task-relevant and task-irrelevant observation factors within each sub-dataset (Figure 6), while maintaining factor independence during data collection (Figure 7).

2. Maintain substantial overlap in the most important factors across sub-datasets (Figure 6), preserving consistency for less critical factors (Section 5).

3. Allow slightly larger distributional disparities for task-relevant factors between sub-datasets, while minimizing disparities in task-irrelevant factors (The last paragraph in Section 3.2).

Furthermore, we give suggestions on how to alleviate the shortcut learning in existing offline datasets, facilitating scenarios where collecting new data is infeasible. We demonstrate that carefully selected robotic data augmentation strategies can effectively increase the diversity within sub-datasets and decrease their differences. Our experiments, conducted in SIMPLER [4] and real-world environment, confirm that these augmentation strategies can significantly alleviate shortcut learning in Generalist Robot Policies like π_0, and improve generalization performance.

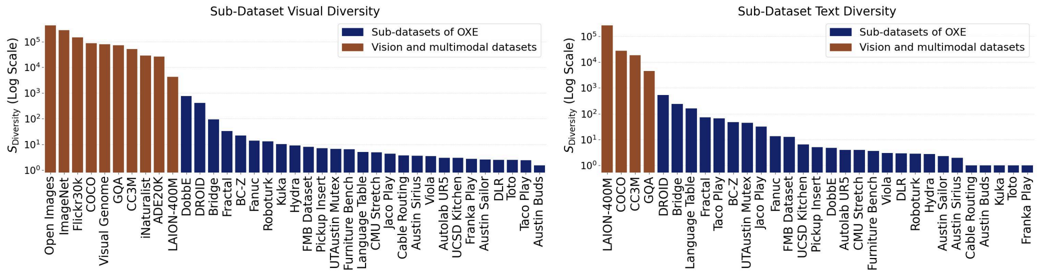


Figure 2: **Comparison of visual (left) and text (right) diversity (log scale) between OXE Sub-Datasets and vision/multimodal Datasets.** OXE sub-datasets exhibit significantly lower diversity compared to their vision and multimodal counterparts. We simply chose $t = 20$ as it does not influence the general trend.

2 Analysis of Dataset Diversity and Fragmentation of Robot Datasets

In this section, before delving into the details of shortcut learning, we analyze the sub-dataset diversity and fragmentation of current large-scale robot datasets. Our discussion centers on the OXE dataset [10], the largest open-source robot dataset utilized for the pretraining of generalist robot policies [10, 13, 14, 7, 16, 20]. Specifically, we focus on OXE Magic Soup++ [13], which comprises 27 sub-datasets from OXE that have been carefully selected to ensure high quality and have been used in several models [13, 14, 7, 16]. Given that current robot datasets for large-scale pretraining similarly consist of diverse sub-datasets [15, 21, 22, 20], the insights derived in this section are broadly applicable.

As most generalist robot policies use vision observations and language instructions for making actions, we utilize visual and language features of the datasets to measure the diversity within sub-datasets and disparities between them, as also suggested by [18]. For visual features, we use the concatenation of features from pretrained DINOv2 [23] and SigLIP [24] as they are shown to give complementary information about images [25, 26]. We focus solely on the initial visual observation of each episode, as subsequent frames typically exhibit minimal variation. Language features are extracted using CLIP [27] for its strong vision-language alignment. To quantify the diversity within a sub-dataset D_i, we adopt the uniformity metric proposed by [28]:

$$S^{D_i}_{\text{diversity}} \triangleq \frac{1}{\mathbb{E}_{u,v\sim D_i}\left[e^{-t\|u-v\|_2^2}\right]},$$

which is maximized when the feature vectors within the sub-dataset are uniformly distributed on the unit sphere [28], indicating maximum diversity (we normalize all features before calculation). This aligns with the entropy measure used in Section 3 to quantify diversity. As illustrated in Figure 10, the temperature parameter t serves as a soft threshold, modulating the influence of pairwise distances $\|u - v\|_2^2$ on the diversity metric. A larger t sharpens the effective influence range, ensuring only highly similar vectors contribute significantly.

Similarly, the disparity metric is defined as the inverse of the expected pairwise similarity between datasets:

$$S_{\text{disparity}} \triangleq \frac{m(m-1)}{\sum_{i\neq j}\mathbb{E}_{u\sim D_i,v\sim D_j}\left[e^{-t\|u-v\|_2^2}\right]},$$

where we assume there are m sub-datasets. In the followings, we derive insights by comparing $S_{\text{diversity}}$ and $S_{\text{disparity}}$ metrics obtained from the OXE dataset with those from commonly used vision and multimodal datasets for pretraining large-scale vision models and vision-language models.

Large-scale robot datasets exhibit limited diversity within individual sub-datasets. As depicted in Figure 2, the visual and textual diversity across all sub-datasets within OXE is markedly lower than that of vision and multimodal datasets. Even the most recent dataset, DROID [9], which aims to improve diversity, remains significantly less diverse by several orders of magnitude. This limited diversity within sub-datasets primarily stems from intrinsic constraints in the collection process of robot datasets. Factors such as scenes and viewpoints are challenging to vary significantly across

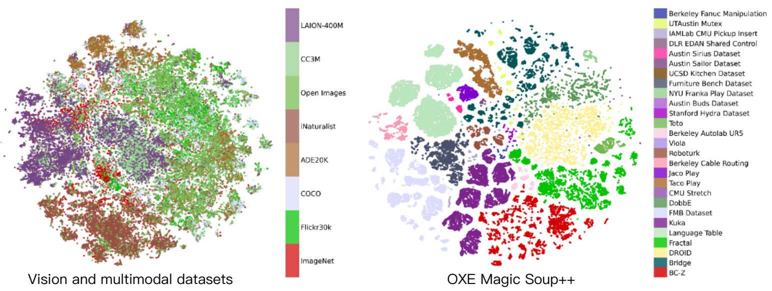


Figure 3: **Comparison of t-SNE visualizations for vision/multimodal datasets (left) versus OXE Magic Soup++ (right).** The figure shows the clear data fragmentation in the OXE dataset, in contrast to the more intertwined data structure observed in the visual and multimodal datasets.

episodes, resulting in a lack of portability compared to web-based vision-language datasets. Another reason is that, as shown in Figure 13, the robotic skills within each sub-dataset are typically predefined, restricting them to a narrow spectrum of tasks.

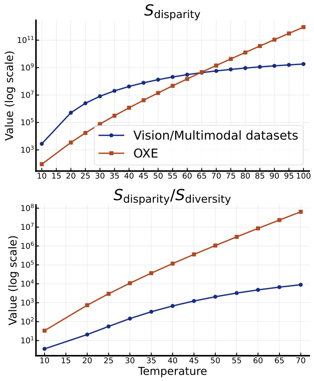


Figure 4: Comparison of the visual disparity metric $S_{\text{disparity}}$ (top) and the combined metric $\frac{S_{\text{disparity}}}{S_{\text{diversity}}}$ (bottom) between OXE and vision/multimodal datasets at different temperatures.

Large-scale robot datasets are fragmented across sub-datasets. We present the visualization of visual features using t-SNE in Figure 3. Unlike vision and multimodal datasets, where different datasets are often intertwined, the sub-datasets of OXE exhibit distinct separations with minimal overlap. Furthermore, some sub-datasets have several separated clusters, effectively fragmenting the whole dataset into more sub-datasets with smaller size. We show examples in Appendix C and further discuss this in Section 3.3.

The top plot in Figure 4 presents the disparity metric $S_{\text{disparity}}$ for OXE. Notably, it is higher than that of vision/multimodal datasets at higher temperatures and lower at lower temperatures. This characteristic is typical of robot datasets: distances between data points from different OXE sub-datasets are concentrated within a specific range. Conversely, distances between data points from different vision and multimodal datasets are more dispersed. Two key factors contribute to this pattern: (1) Robotic scenarios are usually confined to limited in-room tabletop domains, which restricts the maximum possible distances and results in overall high similarity, thus lower disparity at lower temperatures. (2) Fragmentation of data across sub-datasets prevents distances from falling below a certain threshold, establishing a lower bound, which leads to higher disparity when only close data points are considered at higher temperatures.

The bottom plot in Figure 4 illustrates the ratio $\frac{S_{\text{disparity}}}{S_{\text{diversity}}}$, which integrates both diversity and disparity metrics to assess the extent of dataset fragmentation. Both a deficiency in diversity within sub-datasets and an increase in disparity between them result in sub-datasets behaving like isolated "points" scattered across the space, rather than forming a cohesive, interconnected dataset.

Task instructions are distinct but similar across sub-datasets. Despite the lack of overlap in task instructions between sub-datasets, as shown in the left plot of Figure 12 and Table 3, text features from different sub-datasets are closer in space compared to those from vision and multimodal datasets. This similarity arises from shared robotic skills, such as pick-and-place and open/close tasks, and the consistency of text instructions within the same domain.

3 The Role of Dataset Diversity and Fragmentation in Shortcut Learning

In this section, we prove that both of the lack of diversity within sub-datasets and a large disparity (fragmentation) between them lead to shortcut learning. We first describe shortcut learning in detail.

3.1 Shortcut Learning

In the standard supervised or imitation learning framework, we aim to learn a model $\pi_\theta(y|x)$ that maps an observation x to a target y. Following prior work [29], we assume that any observation x is generated from a set of underlying "observation factors." These factors can be divided into two groups: *task-relevant factors* (u), such as object positions or language instructions, which causally determine the target y (i.e., $p(y|x) = p(y|u)$); and *task-irrelevant factors* (v), such as image backgrounds, viewpoints, or robot arm type. *Shortcut learning* is characterized as the scenario where the learned model π_θ improperly relies on these irrelevant factors, meaning its prediction is not conditionally independent of v given u ($\pi_\theta(y|x) \neq \pi_\theta(y|u)$). This critical issue arises when, within the training distribution, the task-relevant and task-irrelevant factors are not independent, i.e., $p_{\text{train}}(u,v) \neq p_{\text{train}}(u)p_{\text{train}}(v)$. This statistical dependency induces a spurious correlation between the irrelevant factor v and the target y, which the model may exploit. Consequently, a model that learns via such shortcuts will exhibit poor performance on out-of-distribution data where these spurious correlations are no longer present.

3.2 The Reasons Behind Shortcut Learning on Robot Data

We first establish a mathematical framework to analyze how correlations can arise in a dataset composed of multiple distinct sub-datasets. Consider a dataset D characterized by two random variables, $u \sim p_u(u)$ and $v \sim p_v(v)$, with supports $U, V \subset \mathbb{R}$. We model D as a mixture of m sub-datasets, $\{D_1, D_2, \ldots, D_m\}$, where each sub-dataset D_i has its own distributions $u_i \sim p_{u_i}(u_i)$ and $v_i \sim p_{v_i}(v_i)$ with supports U_i and V_i. The overall supports are thus $U = \cup_{i=1}^m U_i$ and $V = \cup_{i=1}^m V_i$.

We make the following simplifying assumptions for our analysis:

1. Intra-dataset Independence: Within any given sub-dataset D_i, the variables u_i and v_i are independent, i.e., $p_i(u,v) = p_{u_i}(u)p_{v_i}(v)$.
2. Uniform Mixture: The overall dataset D is a uniform mixture of the sub-datasets, such that $p_u(u) = \frac{1}{m}\sum_{i=1}^m p_{u_i}(u)$ and $p_v(v) = \frac{1}{m}\sum_{i=1}^m p_{v_i}(v)$.

The first assumption is approximately valid, as each sub-dataset is collected under controlled conditions, minimizing the introduction of dependencies between factors. To quantify the correlation between u and v across the entire dataset D, we use the normalized mutual information:

$$\overline{I}(u,v) = \frac{2I(u,v)}{H(u)+H(v)},$$

where $I(u,v)$ is the standard mutual information and $H(\cdot)$ is the Shannon entropy. For simplicity, the following propositions are presented for the case of $m = 2$ sub-datasets.

Proposition 3.1 (Mutual information in disjoint sets)**.** *Given two sub-datasets where the supports for both variables are disjoint, i.e.,* $U_1 \cap U_2 = \varnothing$ *and* $V_1 \cap V_2 = \varnothing$*, the normalized mutual information between* u *and* v *is given by:*

$$\overline{I}(u,v) = \frac{4}{C_{\text{diversity}}+4}, \tag{1}$$

where $C_{\text{diversity}} = H(u_1) + H(u_2) + H(v_1) + H(v_2)$ *is the sum of entropies.*

Proposition 3.2 (Mutual information in overlapping sets)**.** *Given two sub-datasets with potentially overlapping supports, let* $U_{12} = U_1 \cap U_2$ *and* $V_{12} = V_1 \cap V_2$*. The normalized mutual information is bounded by:*

$$\overline{I}(u,v) \leq 1 - \frac{C_{\text{diversity}}}{C_{\text{diversity}} + (4 - C_{\text{interleave}})}, \tag{2}$$

where $C_{\text{interleave}} = \sum_{u \in U_{12}} [p_{u_1}(u) + p_{u_2}(u)] + \sum_{v \in V_{12}} [p_{v_1}(v) + p_{v_2}(v)]$ *quantifies the degree of overlap (interleaving) between the sub-datasets.*

The propositions above provide a formal basis for our core claims, resting on two key assumptions: first, we model task-relevant and task-irrelevant factors as variables u and v, respectively. Second, we assume that large, multi-source datasets like OXE can be approximated by our mixture model.

Lack of diversity strengthens spurious correlations. Proposition 3.1 mathematically demonstrates this intuition. It shows that when sub-datasets are highly fragmented (disjoint supports), the mutual information (our proxy for spurious correlation) is inversely proportional to $C_{\text{diversity}}$ (our proxy for the total diversity within sub-datasets). A robotic model trained on such a dataset can easily learn to associate a task-irrelevant factor (e.g., a specific viewpoint) with a particular sub-dataset, which in turn reveals information about the task-relevant factor, creating a shortcut.

Interleaving sub-datasets weakens spurious correlations. Proposition 3.2 provides theoretical support for this claim. It shows that as the degree of interleaving ($C_{\text{interleave}}$) increases, the upper bound on the mutual information tightens and moves towards zero. Intuitively, when sub-datasets share common factors (e.g., the same objects appear from multiple viewpoints), it becomes impossible for the model to use those factors as reliable shortcuts to identify the sub-dataset of origin, forcing it to learn the true causal relationships.

Impact of the disparity between non-overlapping sub-datasets on shortcut learning. While Proposition 3.2 elucidates how sub-dataset intersections affect shortcut learning, the influence of distance between non-overlapping sub-datasets remains less clear. As shown in Section 2, the OXE dataset exhibits minimal interleaving of visual and textual features across sub-datasets, yet the textual feature distance (u) is notably smaller than the visual distance (v). We hypothesize that larger sub-dataset distances in task-irrelevant features exacerbate shortcut learning. This stems from two key observations: (1) neural networks prioritize learning simpler patterns first [30, 31], and (2) larger feature distances imply greater variance. When task-irrelevant features have substantially greater between-sub-dataset distances than task-relevant ones, models preferentially learn these higher-variance features, forming shortcuts. In OXE, this explains the model's tendency to rely on visual cues over text instructions (Figure 1). We formalize this intuition through gradient analysis of linear models in Appendix J.

3.3 Experimental Verification on LIBERO

To empirically validate our theoretical claims that low intra-dataset diversity and high inter-dataset disparity foster shortcut learning, we conduct controlled experiments on the LIBERO-Spatial task suite [32]. In this setup, featuring a simulated Franka Emika Panda arm with demonstrations containing camera images and language instructions, we define the task-relevant factors (u) as the object's position and the corresponding language instruction. The camera viewpoint serves as the primary task-irrelevant factor (v), mirroring the significant viewpoint variations observed across sub-datasets in large-scale robot datasets like OXE. We also include a real-world experimental verification in Appendix D.

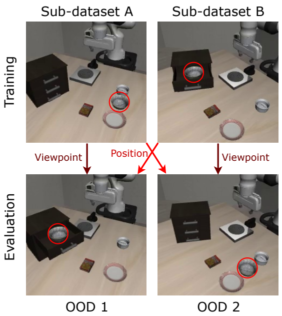


Figure 5: An example of our LIBERO experiment setting, with only one task (or equivalently, one object position/language) within each sub-dataset.

To quantify shortcut learning, we construct a training dataset with a strong spurious correlation between a task-irrelevant viewpoint and a task-relevant object position, and then evaluate the trained policy on out-of-distribution (OOD) configurations where this correlation is broken. We systematically vary the properties of the training data, such as *viewpoint diversity* (the radius of the viewpoint range) and *viewpoint disparity* (the distance between viewpoint centers), to analyze their impact. Performance is measured by two key metrics: (1) the *OOD success rate*, which directly measures generalization, and (2) the *degree of shortcut learning*, a human-assessed score quantifying the model's reliance on the spurious viewpoint cue. Further experimental details are available in Appendix E.

Models. We evaluate three models: (1) **Diffusion Policy** [33], a purely visual policy without language input, utilizing a ResNet-18 architecture; (2) **MiniVLA** [34], a VLA with the same autoregressive structure as OpenVLA [14], but with fewer than 1 billion parameters; (3) $\boldsymbol{\pi_0}$ [7], a strong VLA employing a flow matching objective, pretrained on large-scale robot datasets. While Diffu-

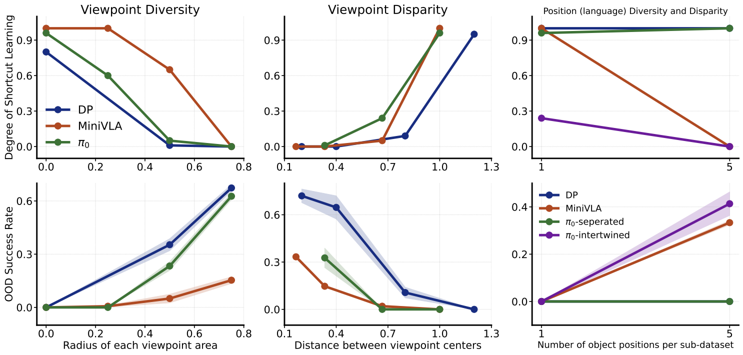


Figure 6: Impact of sub-dataset diversity and disparity on the degree of shortcut leaning and out-of-distribution (OOD) performance of robot policies, analyzing task-relevant factors (object position, language) and task-irrelevant factors (viewpoint). **Note**: Performance metrics are not directly comparable across models due to intentionally varied experimental settings (see Appendix E.4).

sion Policy and MiniVLA are trained from scratch, we finetune π_0 from the pretrained checkpoints with LoRA in order to investigate whether pretrained models are still prone to shortcut learning.

Results. As shown in Figure 6, enhancing diversity within sub-datasets and minimizing disparity between them effectively reduces shortcut dependencies across all evaluated models, aligning with our theoretical analysis. This improvement holds for both task-irrelevant (e.g., viewpoint) and task-relevant factors (e.g., object positions and language variations). Notably, when diversity is increased or disparity decreased, all models transition from complete shortcut reliance (zero success rate) to shortcut-free performance (nonzero success rates). We also note that, increasing object position diversity does not mitigate shortcut learning in the diffusion policy, likely due to the absence of language input. This suggests that without linguistic cues, the model struggles to identify task-relevant features from visual observation alone, underscoring the importance of language instructions.

Diversity does not always help. Previous results are obtained under the assumption of independence of factors within sub-datasets. When diversity breaks factor independence within sub-datasets (e.g., some sub-datasets of OXE; see Figure 11), fragmentation worsens. As illustrated in Figure 7, increasing viewpoint diversity from 2 to 10—while assigning distinct viewpoints to individual tasks—introduces shortcuts and drops OOD success of MiniVLA to zero. Here, viewpoint diversity fragments the original sub-datasets into 10 disjoint subsets, exacerbating fragmentation. **This underscores the need for controlled diversity that preserves factor independence and avoids sub-dataset fragmentation during data collection.**

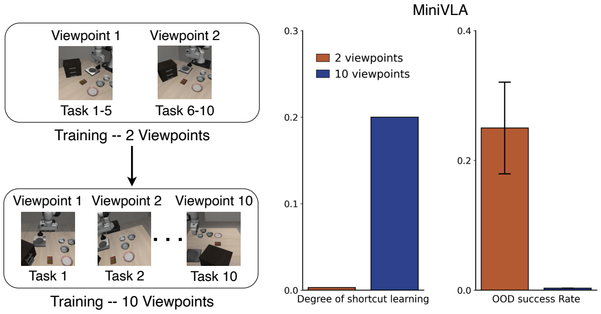


Figure 7: **Diversity does not always help.** Increasing viewpoint diversity by assigning each task a distinct viewpoint induces factor correlations in sub-datasets, aggregating fragmentation.

4 Alleviating Shortcut Learning in Offline Datasets via Data Augmentation

Given that collecting large-scale, perfectly balanced robot datasets from scratch is often prohibitively expensive, a practical alternative is to improve existing offline datasets. In this section, we investigate whether targeted data augmentation strategies can effectively increase sub-dataset diversity and decrease distributional disparities, thereby mitigating shortcut learning.

4.1 Viewpoint Augmentation to Bridge Visual Gaps

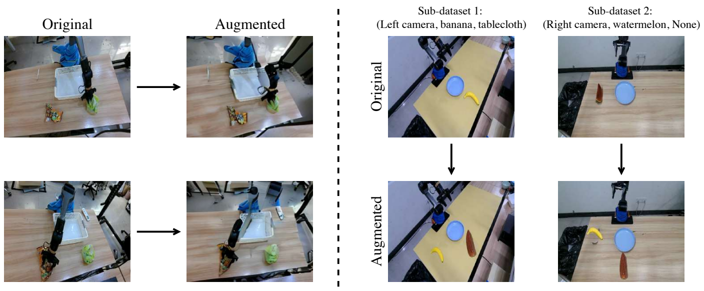

Original
Augmented
Sub-dataset 1:
(Left camera, banana, tablecloth)
Sub-dataset 2:
(Right camera, watermelon, None)
Original
Augmented

Figure 8: Demonstrations of data augmentations used in Section 4. Left: Viewpoint augmentation from one viewpoint to another. Right: Object augmentation that swaps the positions of objects.

Model	Shortcut degree ↓	OOD success rate ↑
π_0 baseline	0.6	0.2
+ viewpoint aug	0.15	0.55

Table 1: Viewpoint augmentation experiment.

Spurious correlations between camera viewpoints and specific tasks are a usual cause of shortcut learning. To address this, we explore the use of viewpoint augmentation methods [35, 36]. In our π_0 fine-tuning experiment (Setup detailed in Section D), we synthetically expand each sub-dataset by generating images from the other's perspective. Specifically, we employ the ZeroNVS model [37] to augment viewpoint A to B and vice-versa for every image, as illustrated in Figure 8. This process effectively breaks the fragmentation of viewpoint factors across the sub-datasets. As evidenced by the results in Table 1, fine-tuning with viewpoint-augmented data significantly reduces the degree of shortcut learning in π_0 and leads to a higher OOD success rate.

4.2 Object Augmentation to Unify Task Distributions

Model	Shortcut degree	
	SIMPLER	Real-world
π_0	1.0	0.8
π_0 + aug	0.68	0.25

Table 2: Comparisons between π_0 with and without object augmentations in the SIMPLER and real-world environment.

To address shortcuts arising from sub-datasets organized around distinct target objects, we employ existing object augmentation techniques [38, 8, 39]. By programmatically swapping objects between different scenes, this method effectively intertwines the object and language distributions, thus reducing inter-dataset disparities (Figure 8). We validated this by fine-tuning a pretrained π_0 model in both SIMPLER [4] and real-world environment (details in Appendix G). As shown in Table 2, the results demonstrate a significant reduction in shortcut behavior. In contrast to the baseline model, which completely fails to follow language instructions in OOD settings, the augmented version exhibits substantially improved language-following and object-reaching capabilities.

5 Discussion and Conclusion

Our analysis reveals that the limited diversity and severe fragmentation in large-scale robot datasets like OXE inherently promote shortcut learning, making naive data scaling detrimental to generalization. This conclusion is supported by the data curation strategies of recent state-of-the-art policies. These models achieve strong performance by either heavily filtering and re-weighting OXE's fragmented sub-datasets [40] or by abandoning it entirely in favor of meticulously controlled data collections where some factors are fixed while others are systematically varied [41, 17].

The overarching takeaway is that pursuing generalization across all factors simultaneously is currently untenable. **The most effective path forward is to strategically control the data collection process: fix non-essential or difficult-to-vary factors while systematically diversifying those of interest**. This disciplined approach is crucial for preventing shortcut learning and provides a clear, actionable framework for training the next generation of robust generalist policies.

6 Limitations and Future Work

While this work provides critical insights into shortcut learning in generalist robot policies, we acknowledge several limitations that open valuable avenues for future research.

Identifying specific shortcuts in large-scale datasets Although our work demonstrates the *existence* of shortcut learning, our analysis does not pinpoint the *specific* spurious correlations exploited by policies trained on massive, heterogeneous datasets like OXE, nor have we investigated the hierarchy of these shortcuts. This limitation points to a clear direction for future work, which should focus on developing fine-grained diagnostic tools and interpretability methods to automatically identify the precise features that models rely on. Such research could involve causal analysis or counterfactual evaluation on large datasets to understand which shortcuts are most dominant and how they vary across different model architectures.

Measuring diversity of task-relevant factors Our quantitative analysis of dataset diversity and disparity primarily focused on task-irrelevant visual features. Due to the significant challenges of collecting and annotating large-scale behavioral data, we could not precisely measure the diversity of *task-relevant* factors, such as the distribution of target object positions or grasp affordances. To address this, we encourage the development of more sophisticated metrics that can capture the complexity of action-centric and object-centric diversity. Exploring semi-supervised or self-supervised methods to automatically label these task-relevant factors would enable a more complete understanding of data quality. Exploring shortcut learning relevant to other observation modalities like proprioceptions [42] and tactile observations is also a promising future direction.

Scalability and generalization of data augmentation Our experiments successfully show that targeted data augmentations can mitigate shortcut learning on a controlled scale, but we have not demonstrated their effectiveness on extremely large datasets like the full OXE collection. The computational cost and potential for introducing artifacts with current augmentation models remain significant challenges at scale. Therefore, a crucial next step is to develop highly efficient, robust, and automated data augmentation pipelines suitable for millions of trajectories. Future work could also systematically compare different augmentation strategies to create a practical guide on the best trade-off between computational cost and performance gain.

Real-world complexity Although we validated our findings in both simulation and real-world setups, the scale and complexity of our real-world experiments are inherently limited and may not fully capture all potential failure modes. Consequently, more extensive real-world studies are needed to validate our findings across a wider range of physical robots, environments, and tasks, including long-term deployments to observe if new, unforeseen shortcuts emerge over time.

Exploring model-centric solutions Our proposed solutions are primarily data-centric, focusing on improving the dataset itself. We did not explore model-centric approaches, such as how different model architectures, training objectives, or regularization techniques might inherently resist shortcut learning, even when trained on fragmented data. A promising future direction is to conduct a comparative analysis of how different model architectures and learning paradigms interact with dataset biases. Investigating hybrid approaches that combine data-centric enhancements with model-centric regularization techniques could lead to the most robust and generalizable robot policies.

Acknowledgments

We would like to thank all reviewers for constructive comments that help us improve the paper. This study is supported by grants from the National Natural Science Foundation of China (Grant No. 62425208, No. U22A2097, No. 62122018, No. 62020106008, No. U23A20315), and Fundamental Research Funds for the Central Universities (Grant No. 22120240634).

References

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[2] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[3] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

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

[5] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

[6] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[7] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[8] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4788–4795. IEEE, 2024.

[9] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[10] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[11] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Huang, S. Jiang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[12] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[13] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

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

[15] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

[16] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

[17] G. R. Team, S. Abeyruwan, J. Ainslie, J.-B. Alayrac, M. G. Arenas, T. Armstrong, A. Balakrishna, R. Baruch, M. Bauza, M. Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[18] J. Gao, S. Belkhale, S. Dasari, A. Balakrishna, D. Shah, and D. Sadigh. A taxonomy for evaluating generalist robot policies. *arXiv preprint arXiv:2503.01238*, 2025.

[19] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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

[21] K. Wu, C. Hou, J. Liu, Z. Che, X. Ju, Z. Yang, M. Li, Y. Zhao, Z. Xu, G. Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024.

[22] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

[23] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[24] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[25] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024.

[26] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*, 2024.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[28] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

[29] I. Higgins, A. Pal, A. Rusu, L. Matthey, C. Burgess, A. Pritzel, M. Botvinick, C. Blundell, and A. Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. In *International conference on machine learning*, pages 1480–1490. PMLR, 2017.

[30] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. C. Courville, Y. Bengio, and S. Lacoste-Julien. A closer look at memorization in deep networks. In *ICML*, pages 233–242. PMLR, 2017.

[31] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. A. Hamprecht, Y. Bengio, and A. C. Courville. On the spectral bias of neural networks. In *ICML*, pages 5301–5310. PMLR, 2019.

[32] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

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

[34] S. Belkhale and D. Sadigh. Minivla: A better vla with a smaller footprint, 2024. URL https://github.com/Stanford-ILIAD/openvla-mini.

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

[36] L. Y. Chen, C. Xu, K. Dharmarajan, M. Z. Irshad, R. Cheng, K. Keutzer, M. Tomizuka, Q. Vuong, and K. Goldberg. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. *arXiv preprint arXiv:2409.03403*, 2024.

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

[38] T. Yu, T. Xiao, A. Stone, J. Tompson, A. Brohan, S. Wang, J. Singh, C. Tan, J. Peralta, B. Ichter, et al. Scaling robot learning with semantically imagined experience. *arXiv preprint arXiv:2302.11550*, 2023.

[39] Z. Chen, Z. Mandi, H. Bharadhwaj, M. Sharma, S. Song, A. Gupta, and V. Kumar. Semantically controllable augmentations for generalizable robot learning. *The International Journal of Robotics Research*, page 02783649241273686, 2024.

[40] J. Hejna, C. Bhateja, Y. Jiang, K. Pertsch, and D. Sadigh. Re-mix: Optimizing data mixtures for large scale imitation learning. *arXiv preprint arXiv:2408.14037*, 2024.

[41] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[42] F. Kuang, J. You, Y. Hu, T. Zhang, C. Wen, and Y. Gao. Adapt your body: Mitigating proprioception shifts in imitation learning. *arXiv preprint arXiv:2506.23944*, 2025.

[43] R. Doshi, H. Walke, O. Mees, S. Dasari, and S. Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. *arXiv preprint arXiv:2408.11812*, 2024.

[44] L. Wang, X. Chen, J. Zhao, and K. He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *Advances in Neural Information Processing Systems*, 37: 124420–124450, 2024.

[45] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, pages 879–893. PMLR, 2018.

[46] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

[47] A.-C. Cheng, Y. Ji, Z. Yang, Z. Gongye, X. Zou, J. Kautz, E. Bıyık, H. Yin, S. Liu, and X. Wang. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453*, 2024.

[48] S. Belkhale, T. Ding, T. Xiao, P. Sermanet, Q. Vuong, J. Tompson, Y. Chebotar, D. Dwibedi, and D. Sadigh. Rt-h: Action hierarchies using language. *arXiv preprint arXiv:2403.01823*, 2024.

[49] L. X. Shi, B. Ichter, M. Equi, L. Ke, K. Pertsch, Q. Vuong, J. Tanner, A. Walling, H. Wang, N. Fusai, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025.

[50] H. Huang, F. Liu, L. Fu, T. Wu, M. Mukadam, J. Malik, K. Goldberg, and P. Abbeel. Otter: A vision-language-action model with text-aware visual feature extraction. *arXiv preprint arXiv:2503.03734*, 2025.

[51] M. J. Kim, C. Finn, and P. Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

[52] H.-S. Fang, H. Fang, Z. Tang, J. Liu, C. Wang, J. Wang, H. Zhu, and C. Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. *arXiv preprint arXiv:2307.00595*, 2023.

[53] Z. Jiang, Y. Xie, K. Lin, Z. Xu, W. Wan, A. Mandlekar, L. Fan, and Y. Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. *arXiv preprint arXiv:2410.24185*, 2024.

[54] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[55] P. Izmailov, P. Kirichenko, N. Gruver, and A. G. Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, 2022.

[56] W. Ye, G. Zheng, X. Cao, Y. Ma, and A. Zhang. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*, 2024.

[57] K. Xiao, L. Engstrom, A. Ilyas, and A. Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

[58] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[59] X. Luo, L. Wei, L. Wen, J. Yang, L. Xie, Z. Xu, and Q. Tian. Rectifying the shortcut learning of background for few-shot learning. *Advances in Neural Information Processing Systems*, 34:13073–13085, 2021.

[60] M. Moayeri, P. Pope, Y. Balaji, and S. Feizi. A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19087–19097, 2022.

[61] A. Kolesnikov and C. H. Lampert. Improving weakly-supervised object localization by micro-annotation. *arXiv preprint arXiv:1605.05538*, 2016.

[62] A. Rosenfeld, R. Zemel, and J. K. Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

[63] S. Singla and S. Feizi. Salient imagenet: How to discover spurious features in deep learning? *arXiv preprint arXiv:2110.04301*, 2021.

[64] R. Shetty, B. Schiele, and M. Fritz. Not using the car to see the sidewalk–quantifying and controlling the effects of context in classification and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8218–8226, 2019.

[65] M. A. Alcorn, Q. Li, Z. Gong, C. Wang, L. Mai, W.-S. Ku, and A. Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4845–4854, 2019.

[66] S. Mo, H. Kang, K. Sohn, C.-L. Li, and J. Shin. Object-aware contrastive learning for debiased scene representation. *Advances in Neural Information Processing Systems*, 34:12251–12264, 2021.

[67] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

[68] W. Brendel and M. Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

[69] Y. Li, Y. Li, and N. Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018.

[70] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*, 2020.

[71] Y. Yuan, L. Zhao, K. Zhang, G. Zheng, and Q. Liu. Do llms overcome shortcut learning? an evaluation of shortcut challenges in large language models. *arXiv preprint arXiv:2410.13343*, 2024.

[72] R. Tang, D. Kong, L. Huang, and H. Xue. Large language models can be lazy learners: Analyze shortcuts in in-context learning. *arXiv preprint arXiv:2305.17256*, 2023.

[73] M. Du, F. He, N. Zou, D. Tao, and X. Hu. Shortcut learning of large language models in natural language understanding. *Communications of the ACM*, 67(1):110–120, 2023.

[74] W. Ding, L. Shi, Y. Chi, and D. Zhao. Seeing is not believing: Robust reinforcement learning against spurious correlation. *Advances in Neural Information Processing Systems*, 36:66328–66363, 2023.

[75] N. Grandien, Q. Delfosse, and K. Kersting. Interpretable end-to-end neurosymbolic reinforcement learning agents. *arXiv preprint arXiv:2410.14371*, 2024.

[76] D. Hoftijzer, G. Burghouts, and L. Spreeuwers. Language-based augmentation to address shortcut learning in object-goal navigation. In *2023 Seventh IEEE International Conference on Robotic Computing (IRC)*, pages 1–8. IEEE, 2023.

[77] Z. Deng, J. Jiang, G. Long, and C. Zhang. Causal reinforcement learning: A survey. *arXiv preprint arXiv:2307.01452*, 2023.

[78] R. Tian, C. Xu, M. Tomizuka, J. Malik, and A. Bajcsy. What matters to you? towards visual representation alignment for robot learning. *arXiv preprint arXiv:2310.07932*, 2023.

[79] R. R. Sanchez, H. Nemlekar, S. Sagheb, C. M. Nunez, and D. P. Losey. Recon: Reducing causal confusion with human-placed markers. *arXiv preprint arXiv:2409.13607*, 2024.

[80] P. De Haan, D. Jayaraman, and S. Levine. Causal confusion in imitation learning. *Advances in neural information processing systems*, 32, 2019.

[81] J. Park, Y. Seo, C. Liu, L. Zhao, T. Qin, J. Shin, and T.-Y. Liu. Object-aware regularization for addressing causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 34:3029–3042, 2021.

[82] I. Bica, D. Jarrett, and M. van der Schaar. Invariant causal imitation learning for generalizable policies. *Advances in Neural Information Processing Systems*, 34:3952–3964, 2021.

[83] Y. Chen, Y. Zhang, G. D'urso, N. Lawrance, and B. Tidd. Improving generalization ability of robotic imitation learning by resolving causal confusion in observations. *arXiv preprint arXiv:2507.22380*, 2025.

[84] J. Zhang, S. Wu, X. Luo, H. Wu, L. Gao, H. T. Shen, and J. Song. Inspire: Vision-language-action models with intrinsic spatial reasoning. *arXiv preprint arXiv:2505.13888*, 2025.

[85] S. Wu, J. Zhang, X. Luo, J. Xie, J. Song, H. T. Shen, and L. Gao. Policy contrastive decoding for robotic foundation models. *arXiv preprint arXiv:2505.13255*, 2025.

[86] J. Hejna, S. Mirchandani, A. Balakrishna, A. Xie, A. Wahid, J. Tompson, P. Sanketi, D. Shah, C. Devin, and D. Sadigh. Robot data curation with mutual information estimators. *arXiv preprint arXiv:2502.08623*, 2025.

[87] S. Bai, W. Zhou, P. Ding, W. Zhao, D. Wang, and B. Chen. Rethinking latent redundancy in behavior cloning: An information bottleneck approach for robot manipulation. *arXiv preprint arXiv:2502.02853*, 2025.

[88] S. Belkhale, Y. Cui, and D. Sadigh. Data quality in imitation learning. *Advances in neural information processing systems*, 36:80375–80395, 2023.

[89] J. Hejna, C. Bhateja, Y. Jiang, K. Pertsch, and D. Sadigh. Re-mix: Optimizing data mixtures for large scale imitation learning. *arXiv preprint arXiv:2408.14037*, 2024.

[90] J. Gao, A. Xie, T. Xiao, C. Finn, and D. Sadigh. Efficient data collection for robotic manipulation via compositional generalization. *arXiv preprint arXiv:2403.05110*, 2024.

[91] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[92] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

[93] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[94] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

[95] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[96] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[97] D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[98] Z. Lin, C. Liu, R. Zhang, P. Gao, L. Qiu, H. Xiao, H. Qiu, C. Lin, W. Shao, K. Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

[99] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

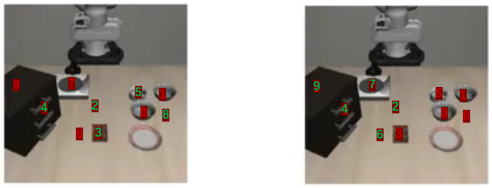
[100] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[101] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

A Related Work

Generalist robot policies. Following the trend in machine learning research, multiple works have developed robotic foundation models [6, 12, 13, 15, 43, 44, 14, 45], in particular Vision-Language-Action (VLA) models [12, 14, 7, 16, 46, 47, 48, 17, 20, 41, 49, 50, 34, 51]. By pretraining on increasingly large robot datasets [10, 52, 5, 9, 8, 53, 11], these models produce generalist robot policies that excel at a wide variety of tasks and exhibit some degree of generalization [12, 13]. However, research by [18] suggests that training on large-scale datasets does not significantly enhance the generalization capabilities of these policies. In particular, current models still struggle to generalize to many environmental changes, including viewpoint, language, object poses, etc. Our work delves into the problem and shows that limited diversity within individual sub-datasets, and significant distributional disparities across sub-datasets lead to shortcut learning of policies, which hinders generalization. Recent VLAs such as π_0 [7], $\pi_{0.5}$ [41], and Gemini Robotics [17] have demonstrated enhanced generalization capabilities by collecting diverse, large-scale datasets within controlled environments, where certain factors such as tasks, scene types, and embodiments are fixed, while others are varied. This mitigates data fragmentation, supporting our theoretical framework.

Shortcut learning in neural networks. Neural networks are known to exploit spurious correlations for decision-making, leading to the shortcut learning of non-robust features or confounding factors, which can significantly hinder generalization [54, 55, 56]. In vision tasks, neural networks have been observed to rely on multiple task-irrelevant factors, including image backgrounds [57, 58, 59, 60], secondary objects [61, 62, 63, 64, 65, 66], object textures [67] and other confounding factors [68, 69]. In the language domain, recent studies have demonstrated that large language models tend to exploit dataset biases as shortcuts for making predictions in various downstream tasks [70, 71, 72, 73]. While there are a few works discussing shortcut learning in reinforcement learning [74, 75, 76, 77, 78] and imitation learning [79, 80, 81, 82, 83], to the best of our knowledge, we are the first to investigate shortcut learning in generalist robot policies developed through imitation learning on large-scale datasets. Building on this work, recent studies have proposed new methods



Spatially intertwined

Spatially seperated

Figure 9: **Experimental setup for object position layouts across 10 tasks.** Objects from the same sub-dataset share the same color. In the left plot, object positions are spatially intertwined between sub-datasets, whereas in the right plot, they are spatially separated. Unless otherwise specified, experiments employ the high-disparity configuration (right).

to mitigate shortcut learning in gereralist robot policies, either by modifying the training data [84] or through training-free approaches [85].

Recent works have also applied information-theoretic concepts to robotics. For instance, Hejna et al. [86] use mutual information estimators to score the quality of individual demonstration trajectories for data curation, focusing on intra-trajectory properties like action diversity and predictability. Separately, Bai et al. [87] apply the Information Bottleneck principle as a regularization technique during training to mitigate redundancy in the model's latent representations. Our work is distinct from both. Rather than using mutual information as a trajectory scoring function or a model regularizer, we employ it as a diagnostic tool at the dataset-structure level. Our analysis reveals how properties between sub-datasets—namely fragmentation and limited diversity—give rise to spurious correlations. We demonstrate that these structural flaws are a root cause of shortcut learning, a problem orthogonal to the quality of individual trajectories or the redundancy of a model's learned representation.

Recent research has increasingly recognized the critical role of data quality in imitation learning, moving beyond simple heuristics like dataset size. For instance, Belkhale et al. [88] provide a formalism for data quality through the lens of distribution shift, identifying key intra-trajectory properties like action divergence and transition diversity as crucial for policy performance. Operating at the level of entire datasets, Hejna et al. [89] tackle the challenge of composing large-scale, heterogeneous data mixtures. Their method, Re-Mix, uses distributionally robust optimization to learn optimal sampling weights for different data domains, demonstrating that the composition of the training data has an outsized impact on the final policy's generalization capabilities. While these works aim to mitigate distribution shift by analyzing trajectory-level properties or by optimizing the dataset mixture, our work addresses the distinct but related problem of *shortcut learning*. Our contribution is a novel analysis at the dataset-structure level. We use information-theoretic principles not as a trajectory scoring function or a mixture optimization objective, but as a diagnostic tool to reveal how structural flaws—namely high fragmentation and low diversity across sub-datasets—are a fundamental cause of the spurious correlations that lead to shortcut behaviors. Thus, our focus is on diagnosing the origin of a specific failure mode rooted in the dataset's structure, rather than on general data curation or mixture optimization.

A recent study [90] also investigates generalization across factors in a compositional manner, similar to the setting we study in Figure 1. However, the primary focus of that work is on optimizing data collection to cover all possible factors, rather than investigating shortcuts or spurious correlations.

B The Influence of Temperature

In Figure 10, we present a visualization of the similarity metric function $e^{-t\|x\|_2^2}$, as discussed in Section 2. This function is examined under varying values of the temperature parameter t. As t increases, the function's value approaches zero more rapidly. Consequently, the temperature t

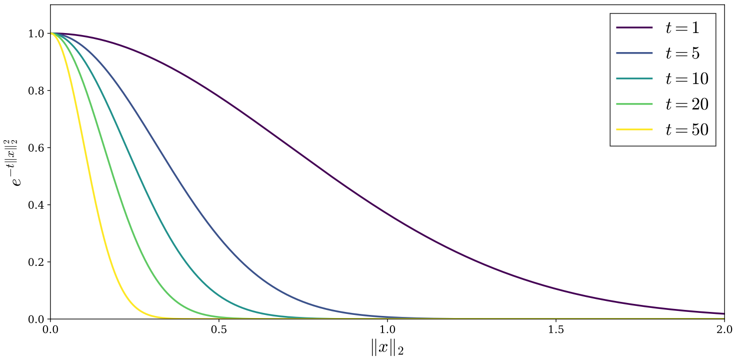


Figure 10: The similarity metric in [28] when varying the temperature t.

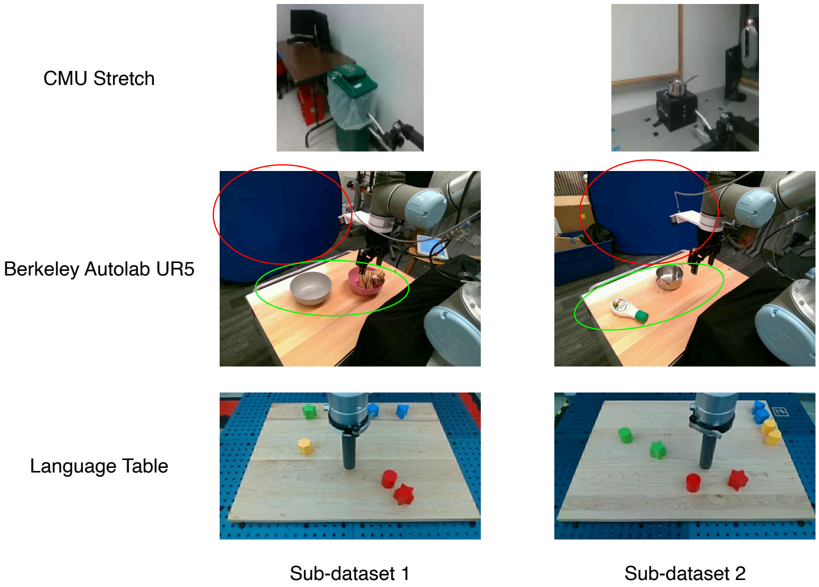


Figure 11: Three fragmented sub-datasets from OXE, each demonstrating distinct fragmentation patterns: (1) CMU Stretch, decomposable into disjoint scenes and tasks; (2) Berkeley Autolab UR5, featuring several factor with time-correlated variations (e.g., background and tasks); (3) Language Table, with only one sparsely changing factor (e.g., lighting).

effectively establishes a soft threshold, which governs the range of $\|x\|_2$ over which the function maintains a value greater than zero.

C Additional Dataset Analysis

Sub-Dataset Fragmentation Analysis Figure 11 illustrates three characteristic fragmentation patterns of sub-datasets in OXE: (1) *Language Table* exhibits natural clustering due to infrequent lighting changes, creating factor-independent subsets without inducing shortcut learning; (2) *Berkeley AutoLab UR5* demonstrates unintended time-correlated variations where task segments coincide with background changes from human activity, creating spurious task-background correlations that promote shortcut learning; (3) *CMU Stretch* (OXE sub-datasets) contains disjoint scenes with simultaneously varying environmental factors and tasks, forming strongly correlated subsets that exacerbate shortcut learning. These patterns highlight how different data collection processes of each sub-dataset can inadvertently create problematic correlations between environmental factors and tasks, aligning with our experiment results in Figure 7.

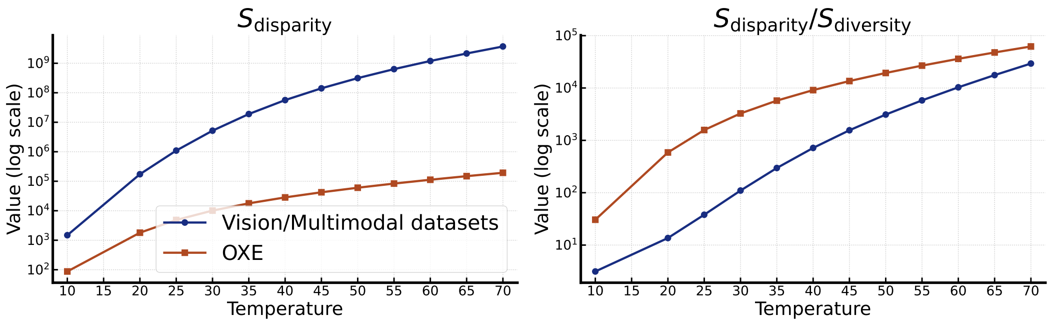


Figure 12: Comparison of the textual disparity metric $S_{\text{disparity}}$ (left) and the combined metric $\frac{S_{\text{disparity}}}{S_{\text{diversity}}}$ (right) between OXE and vision/multimodal datasets at different temperatures.

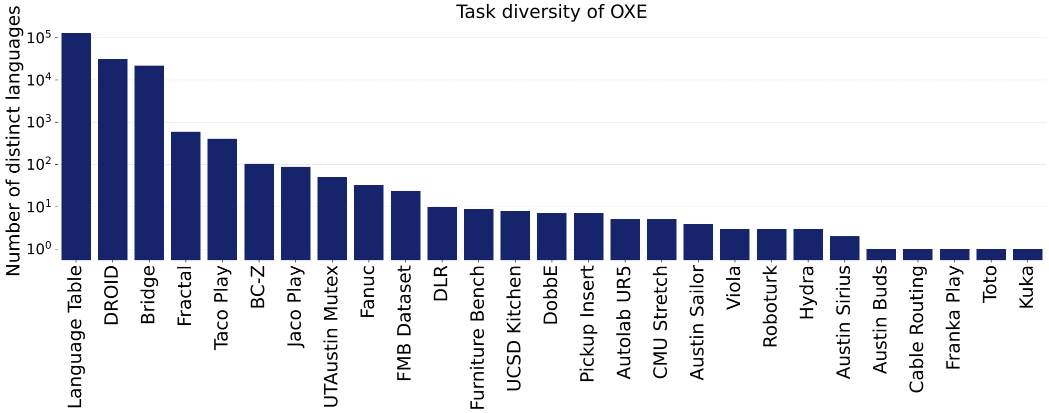


Figure 13: The number of distinct tasks (languages) within each sub-dataset of OXE. Most sub-datasets only have less than 10 tasks, which leads to extremely low task diversity.

D Real-World Experimental Verification

To validate our theoretical conclusions from Section 3.2 in a physical environment, we conducted a real-world experiment. The setup, similar to the one depicted in Figure 1, utilized an AgileX PIPER robotic arm and two cameras positioned at different viewpoints. Initially, we constructed two distinct sub-datasets. Each sub-dataset represented a unique combination of a camera viewpoint (a task-irrelevant factor) and a target object with its corresponding instruction (task-relevant factors). As demonstrated in our preliminary findings (Figure 1), a π_0 model fine-tuned on these two highly-correlated sub-datasets exhibited severe shortcut learning; it learned to associate the viewpoint with the action, ignoring the language instruction.

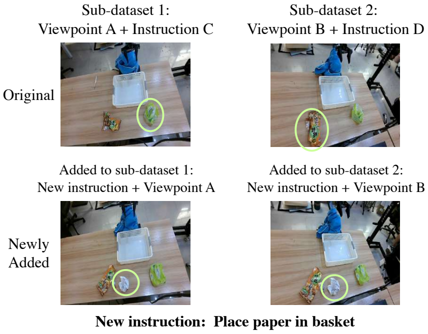

Sub-dataset 1:
Viewpoint A + Instruction C
Sub-dataset 2:
Viewpoint B + Instruction D
Original
Added to sub-dataset 1:
New instruction + Viewpoint A
Added to sub-dataset 2:
New instruction + Viewpoint B
Newly
Added
New instruction: Place paper in basket

Figure 14: Building a "bridge" to connect sub-datasets for the π_0 fine-tuning experiment. Data from a third object is added under both viewpoints.

To investigate how increasing sub-dataset diversity and reducing inter-dataset disparity could mitigate this issue, we introduced new data. Specifically, we added demonstrations involving a third target object, captured from *both* camera viewpoints (as shown in the bottom row of Figure 14). This new data acts as a "bridge" between the original two sub-datasets. By doing so, we simultaneously

Table 3: Sub-dataset task overlap statistics of OXE

Metric	**Value**
Total tasks	182,158
Total overlapping tasks between datasets	165
Percentage of overlapping sub-dataset pairs	3.70 %

increased the instruction diversity within each sub-dataset and decreased the disparity between them, as they now share a common instruction factor.

Model	Shortcut degree ↓	OOD success rate ↑
π_0 baseline	0.6	0.2
+ third object	0	0.75

Table 4: "Third object" experiment.

The results, summarized in Table 4, are compelling. The addition of the third "bridge" object completely eliminated the observed shortcut behavior, leading to a substantial improvement in OOD success rate. By learning from data where the object and instruction were consistent across different viewpoints, the model successfully learned viewpoint invariance. This experiment not only confirms our theoretical framework in a real-world setting but also suggests a valuable strategy for data collection: deliberately creating "bridge" data by varying one factor while keeping others constant can effectively connect disparate sub-datasets, break spurious correlations, and enhance the generalization capabilities of robot policies.

E LIBERO Experiment Details

E.1 Model Implementation and Training

For our analysis, we implemented and trained three distinct models, each with specific configurations:

- **Diffusion Policy:** This model uses a ResNet-18 vision backbone with images resized to 84×84. It was trained for 30,000 iterations (batch size 128) using an AdamW optimizer (learning rate 1×10^{-4}, weight decay 1×10^{-6}). The model takes a 2-step observation history as input, excluding proprioception. Training required approximately 5 hours on a single NVIDIA 3090 GPU.
- **MiniVLA:** This vision-language-action model uses Vector-Quantized action chunks (horizon=8) and was trained without wrist camera images, proprioception, or historical state data. It was optimized for 10,000 steps (batch size 128) with a constant learning rate of 2×10^{-5}. Training was distributed across eight NVIDIA A6000 GPUs and took 5 hours.
- **π_0:** This model integrates a PaliGemma 2B backbone (using LoRA) with a 300M-parameter action expert (action dimension 32, horizon 50). It was trained for 10,000 steps (batch size 32) using AdamW with a cosine decay learning rate schedule (1,000-step warmup to a peak LR of 2.5×10^{-5}, decaying to 2.5×10^{-6} over 30,000 steps). Training took 8 hours across four NVIDIA A6000 GPUs.

E.2 Experimental Environment and Task Setup

Our evaluations were conducted within the **LIBERO-Spatial** suite. The fundamental goal for all 10 manipulation tasks is to **place the target bowl into the red plate**.

- **Task-Relevant Factors**: The 10 distinct tasks are defined by the **initial position of the target bowl** (e.g., in a drawer, on a shelf). The corresponding **language instruction** changes accordingly to reflect this initial position (e.g., "*pick up the bowl in the top drawer and place it on the red plate*").

- **Task-Irrelevant Factor**: We focused on the **camera viewpoint**, defined by $\theta \in [-10°, 90°]$, as the primary task-irrelevant factor.
- **Scene Simplification**: The original LIBERO environment contains two bowls. To better isolate the factors of interest, we **removed one bowl**, leaving only a single target object (marked in red in Figure 5). This simplification also accommodates vision-only models like Diffusion Policy.

Training and Evaluation Protocol. For each experiment, we construct a training dataset composed of two distinct sub-datasets, D_A and D_B. Each sub-dataset is generated to create a strong spurious correlation between the task-irrelevant viewpoint and the task-relevant position. For instance, demonstrations in D_A exclusively pair a specific range of viewpoints (Viewpoint Range A) with a specific set of object positions (Position Set A), while D_B pairs Viewpoint Range B with Position Set B. The model is trained on the combination of D_A and D_B. To quantify shortcut learning, we evaluate the trained policy on out-of-distribution (OOD) configurations where the learned spurious correlations are broken. Specifically, the evaluation consists of two controlled settings: (1) tasking the model with object positions from Set B but from viewpoints within Range A, and (2) the reverse pairing (positions from Set A, viewpoints from Range B). A model relying on the viewpoint shortcut would fail, as it would incorrectly associate the viewpoint with the training-time positions, ignoring the actual object position and instruction.

Experimental Variables and Metrics. We systematically vary the properties of the training data to analyze their impact. *Viewpoint diversity* is the radius of the viewpoint range within each sub-dataset, while *viewpoint disparity* is the distance between the centers of the two viewpoint ranges. To study the effect of task-relevant diversity and disparity, we vary the number of object positions per sub-dataset (from 1 to 5) and their spatial layout (intertwined vs. separated, see Figure 9). Performance is measured by two key metrics: (1) the **OOD success rate**, averaged over the two OOD settings, which directly measures generalization, and (2) the **degree of shortcut learning**, a human-assessed score quantifying the model's tendency to perform the wrong task based on the irrelevant viewpoint cue (lower is better). To ensure fair comparisons, for a given model, the OOD evaluation viewpoint is kept consistent within each experimental set (e.g., a curve in one plot), where we vary data diversity or disparity.

E.3 Data Collection

For each experimental setting shown in Figure 6, we used **200 demonstrations for each task**. These were generated by sampling 4 random viewpoints for each of the 50 base trajectories provided by LIBERO for that task.

E.4 Protocol for Viewpoint Diversity and Disparity Experiments

- **Parameter Selection Strategy**: For both the diversity and disparity experiments, the specific viewpoint centers and radii were not chosen arbitrarily. They were **systematically selected to identify the critical range where the policy's behavior transitions from robust to shortcut-reliant**. This allowed us to precisely map out the model's sensitivity to these dataset properties.
- **Diversity Protocol (Fig. 6 Left)**: We systematically increased the *range* (radius) of viewpoints for each task during training, while holding the *centers* of the viewpoint distributions constant. Evaluation was performed at fixed, out-of-distribution viewpoints to fairly assess generalization.
- **Disparity Protocol (Fig. 6 Middle)**: Conversely, we varied the *distance between the centers* of the viewpoint distributions while keeping their *radius* constant and narrow. To ensure a challenging test, evaluation points were always selected from the boundaries of the opposing task's distribution.

F Real-world Experiment Setup

This section details the setup for the two distinct real-world experiments presented in the paper. Both experiments utilize an AgileX PIPER robotic arm and are observed by two cameras from different viewpoints.

F.1 Experiments in Figure 1

This experiment is designed to test if a model exhibits shortcut learning when task-irrelevant factors are confounded, even when all objects are present during training.

- **Task-Relevant Factors**: The identity of the target object (banana or watermelon) and the corresponding language instruction (”place tissue bag into the plate” or ”place snack bag into the plate”).
- **Task-Irrelevant Factor**: The camera viewpoint (left or right camera).
- **Training Data Setup**: Two sub-datasets were created.
  - **Sub-dataset 1**: The instruction is ”place tissue bag into the plate”, collected exclusively from the **left camera**.
  - **Sub-dataset 2**: The instruction is ”place snack bag into the plate”, collected exclusively from the **right camera**.
- **Data Collection Details**: A key difference from the object augmentation experiment is that during the collection of each demonstration, **both the tissue bag and the snack bag were present on the table**. The only sources of randomness were the minor variations in the orientation of the objects and the slight shifts in their positions and the position of the plate. We collected 20 demonstrations for each sub-dataset.
- **Evaluation**: The model is evaluated on its ability to follow the correct instruction when the viewpoint is swapped (e.g., given the ”place snack bag...” instruction from the left camera’s viewpoint). The model’s ability to follow the instruction was measured over 10 trials for each condition.

F.2 Object Augmentation Experiments

This experiment is designed to create a strong spurious correlation between objects and multiple visual factors (viewpoint and background) and to test if data augmentation can mitigate the resulting shortcut behavior.

- **Task-Relevant Factors**: The identity of the target object (banana or watermelon) and the corresponding language instruction.
- **Task-Irrelevant Factors**: The camera viewpoint, the background scene, and the positions of objects.
- **Training Data Setup (Confounded)**: These irrelevant factors were deliberately confounded with the task-relevant object.
  - **Sub-dataset 1 (”Banana Env”)**: The task ”put banana into the plate” was collected exclusively from the **left camera** with a **yellow tablecloth** background.
  - **Sub-dataset 2 (”Watermelon Env”)**: The task ”put watermelon into the plate” was collected exclusively from the **right camera** with **no tablecloth**.
- **Data Collection Details**: During training data collection for each sub-dataset, only the single relevant object (either the banana or the watermelon) was present. The object was randomly placed either on the plate or to its right. Each sub-dataset consists of 20 demonstrations.

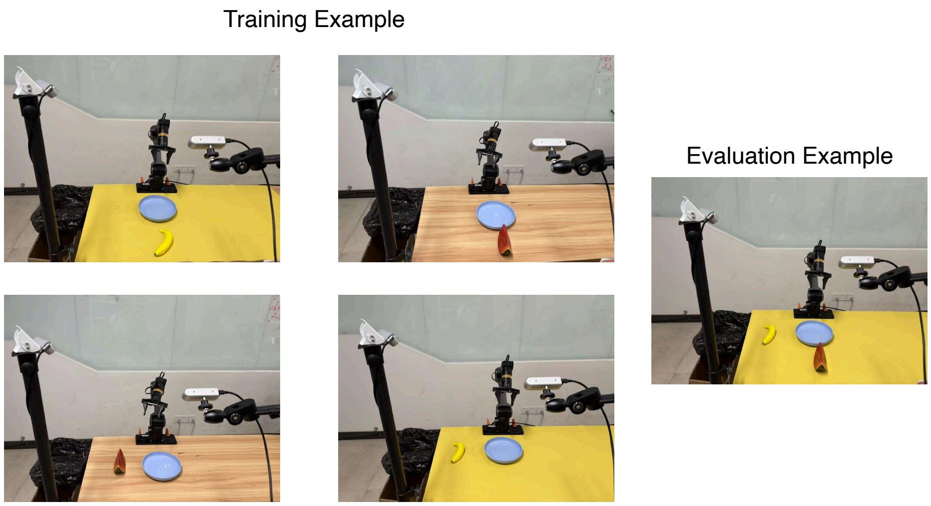


Figure 15: **Real-world object augmentation experiment setup.** Two cameras are positioned in front of an AgileX PIPER robotic arm. For training, two distinct sub-datasets were created, each featuring a single, specific combination of object type (banana or watermelon), camera viewpoint (left or right), and background (with or without a yellow tablecloth). During evaluation, tests were conducted on both two original combinations of viewpoints and backgrounds. In these evaluations, both objects (banana and watermelon) were simultaneously present on the table, and the robot was guided by language instructions referring to object-scene configurations *not* explicitly encountered in the training combinations.

- **Evaluation**: To test for shortcut learning, the learned correlations were disrupted by introducing out-of-distribution (OOD) objects into the scene (e.g., presenting the banana in the right-camera, no-tablecloth environment) and providing the corresponding language instruction. The model's ability to follow the instruction was measured over 10 trials for each condition.

G Object Augmentation Details

G.1 Data Collection and Training Setup

For our experiments involving object augmentation, we established the following datasets and training protocols:

- **SIMPLER Environment:** We collected a total of 242 successful trajectories, comprising 116 from the RT-1 environment and 126 from the Bridge environment.
- **Real-World Environment:** We collected 20 demonstrations for each of the two distinct sub-datasets.
- **Training Protocol:** All models, whether using original or augmented data, were fine-tuned for 2,500 steps to ensure a fair comparison.

G.2 Augmentation Pipeline

Our object augmentation pipeline is designed to decouple objects from their original visual contexts. It consists of the following three stages:

1. **Step 1: Object Mask Library Creation.** First, we build a comprehensive object mask library (D). To do this, we apply Grounded-SAM2 to the initial frame of every episode in our dataset. This allows us to extract high-quality segmentation masks for all target objects as seen from various perspectives, creating a rich library of object assets.

2. **Step 2: Scene Preparation (Object Removal and Inpainting).** Next, for each image (o_t) in a trajectory, we identify the target object specified by the language instruction (L). Using Grounded-SAM2 again, we segment this specific object to obtain its mask (m_{orig}) and record its original position (via the bounding box center, c_{orig}). The object is then digitally erased from the image, and an inpainting model seamlessly reconstructs the background where the object was. This step yields a ”clean” background image, ready for augmentation.

3. **Step 3: Object Swapping and Augmented Dataset Generation.** Finally, we create the augmented image. For a clean scene that originally contained a specific object (e.g., a banana with mask d_{orig}), we randomly sample a mask of a *different* object from our library (e.g., a watermelon, $d_{aug} \in D \setminus \{d_{orig}\}$). We then paste this new object into the clean scene, carefully aligning its center with the original object's location (c_{orig}). To create a more challenging OOD scenario, we then re-introduce the original object into the same scene to act as a distractor. The placement of this distractor depends on the environment: in SIMPLER, it is placed at a random valid location, while in our real-world setup, it is placed in one of the two predefined object locations. By applying this full procedure—swapping the target and adding a distractor—across all images, we generate the final augmented dataset where objects are fully decoupled from their contexts.

H Methodology for Human-Assisted Shortcut Scoring

To quantitatively measure the degree to which a policy relies on shortcut learning during out-of-distribution (OOD) evaluations, we developed a human-assisted scoring methodology. This approach allows for a nuanced assessment of the robot's behavior beyond simple binary success/failure metrics.

The scoring process is conducted as follows:

1. **Video Review:** Human evaluators are presented with video recordings of every evaluation trial for a given experimental setup. Each video captures the complete sequence of actions taken by the robot policy from the start to the end of an episode.

2. **Behavioral Judgment:** For each video, the evaluator judges whether the policy's actions correspond to the given language instruction or if they revert to a “shortcut” behavior learned from spurious correlations in the training data.

3. **Scoring Rubric:** A score is assigned to each trial based on a predefined rubric:

   - **Score = 1.0 (Clear Shortcut):** The policy unequivocally ignores the instruction and performs a clear shortcut action. For example, when instructed to interact with object A from a viewpoint previously associated with object B, the policy ignores A and attempts to interact with B.
   - **Score = 0.5 (Ambiguous or Partial Shortcut):** The policy's behavior is unclear or appears to be a mix of the correct and shortcut actions. This includes cases where the robot targets a location midway between the correct object (A) and the shortcut object (B), or exhibits significant hesitation.
   - **Score = 0.0 (No Shortcut):** The policy correctly attempts to follow the language instruction and does not exhibit any observable shortcut behavior, regardless of whether the attempt is successful or not.

4. **Final Score Calculation:** The final “Degree of Shortcut Learning” for a model is calculated by averaging the scores from all of its evaluation trials. A score closer to 1.0 indicates a strong tendency to rely on shortcuts, while a score closer to 0.0 indicates that the policy is more robust to the spurious correlations present in the training data.

I Proofs of Propositions in Section 3

Proof of Proposition 3.1:

Proof. Given the condition $p_u(u) = \frac{1}{2}\left[p_{u_1}(u) + p_{u_2}(u)\right]$ and $H(u) = -\sum_{u \in U} p_u(u) \log p_u(u)$, we have

$$\begin{aligned} H(u) &= -\sum_{u \in U_1} \frac{p_{u_1}(u)}{2} \log \frac{p_{u_1}(u)}{2} - \sum_{u \in U_2} \frac{p_{u_2}(u)}{2} \log \frac{p_{u_2}(u)}{2} \\ &= \frac{H(u_1) + H(u_2)}{2} + 1. \end{aligned}$$

The last equation comes from the fact that $\sum_{u \in U_1} p_{u_1}(u) \log \frac{p_{u_1}(u)}{2} = H(u_1) + \log 2$ and $log_2 2 = 1$. Similarly, $H(v) = \frac{H(v_1)+H(v_2)}{2} + 1$. For the mutual information, since the assumption of independence of factors within each sub-dataset, we have $I(u_1, v_1) = I(u_2, v_2) = 0$, and thus

$$\begin{aligned} I(u, v) = &\sum_{u \in U_1, v \in V_1} \frac{p_1(u, v)}{2} \log \frac{\frac{p_1(u,v)}{2}}{\frac{p_{u_1}(u)}{2} \cdot \frac{p_{v_1}(v)}{2}} + \sum_{u \in U_2, v \in V_2} \frac{p_2(u, v)}{2} \log \frac{\frac{p_2(u,v)}{2}}{\frac{p_{u_2}(u)}{2} \cdot \frac{p_y(v)}{2}} \\ &= \frac{I(u_1, v_1) + I(u_2, v_2)}{2} + 1 \\ &= 1. \end{aligned}$$

Put together $H(u)$, $H(v)$ and $I(u, v)$, we have

$$\begin{aligned} \bar{I}(u, v) &= \frac{2I(u, v)}{H(u) + H(v)} \\ &= \frac{4}{H(u_1) + H(u_2) + H(v_1) + H(v_2) + 4} \\ &= \frac{4}{C_{\text{diversity}} + 4}, \end{aligned}$$

which completes the proof. □

Proof of Proposition 3.2:

Proof. Since both sub-datasets involve probabilities over U_{12} and V_{12}, we should consider each region separately. First, we calculate the entropy $H(u)$:

$$\begin{aligned} H(u) = &-\sum_{u \in U_1 \setminus U_{12}} \frac{p_{u_1}(u)}{2} \log \frac{p_{u_1}(u)}{2} - \sum_{u \in U_2 \setminus U_{12}} \frac{p_{u_2}(u)}{2} \log \frac{p_{u_2}(u)}{2} \\ &- \sum_{u \in U_{12}} \frac{p_{u_1}(u) + p_{u_2}(u)}{2} \log \frac{p_{u_1}(u) + p_{u_2}(u)}{2}. \end{aligned}$$

By applying Jensen Inequality, we have

$$\frac{p_{u_1}(u) + p_{u_2}(u)}{2} \log \frac{p_{u_1}(u) + p_{u_2}(u)}{2} \leq \frac{p_{u_1}(u)}{2} \log p_{u_1}(u) + \frac{p_{u_2}(u)}{2} \log p_{u_2}(u),$$

which gives

$$\begin{aligned}H(u) \geq & -\sum_{u\in U_1\setminus U_{12}} \frac{p_{u_1}(u)}{2}\log\frac{p_{u_1}(u)}{2} - \sum_{u\in U_2\setminus U_{12}} \frac{p_{u_2}(u)}{2}\log\frac{p_{u_2}(u)}{2} \\ & -\sum_{u\in U_{12}}\left[\frac{p_{u_1}(u)}{2}\log p_{u_1}(u) + \frac{p_{u_2}(u)}{2}\log p_{u_2}(u)\right] \\ =& \frac{1}{2}\left[H(u_1)+H(u_2)+\sum_{u\in U_1\setminus U_{12}} p(u_1) + \sum_{u\in U_2\setminus U_{12}} p(u_2)\right] \\ =& \frac{1}{2}\left[H(u_1)+H(u_2)+2-\sum_{u\in U_{12}} p(u_1) - \sum_{u\in U_{12}} p(u_2)\right].\end{aligned}$$

Similarly, for $H(v)$, we have

$$H(v) \geq \frac{1}{2}\left[H(v_1)+H(v_2)+2-\sum_{v\in V_{12}} p(v_1) - \sum_{v\in V_{12}} p(v_2)\right].$$

Given that $C_{\text{interleave}} = \sum_{u\in U_{12}}[p_{u_1}(u)+p_{u_2}(u)] + \sum_{v\in V_{12}}[p_{v_1}(v)+p_{v_2}(v)]$, we have

$$H(u)+H(v) \geq \frac{1}{2}\left[C_{\text{diversity}} + 4 - C_{\text{interleave}})\right].$$

Then we calculate the mutual information. We partition the calculation into four terms:

$$\begin{aligned}I(u,v) =& \sum_{i=1}^{2}\left(\sum_{u\in U_i\setminus U_{12}, v\in V_i\setminus V_{12}} \frac{p_i(u,v)}{2}\log\frac{2p_i(u,v)}{p_{u_i}(u)p_i(v)}\right) \\ & + \sum_{u\in U_{12}, v\in V} p(u,v)\log\frac{p(u,v)}{p_u(u)p_v(v)} + \sum_{u\in U, v\in V_{12}} p(u,v)\log\frac{p(u,v)}{p_u(u)p_v(v)}.\end{aligned}$$

We first calculate the last two terms. Note that

$$\begin{aligned}& \sum_{u\in U_{12}, v\in V_1\cup V_{12}} p(u,v)\log\frac{p(u,v)}{p_u(u)p_v(v)} \\ =& \sum_{u\in U_{12}, v\in V_1\cup V_{12}} \frac{1}{4}\left[p_1(u,v)(1+\frac{p_{u_2}(u)}{p_{u_1}(u)})\right]\log\frac{p_1(u,v)}{p_{u_1}(u)p_{v_1}(v)}.\end{aligned}$$

Since u and v are independent in the first sub-dataset, we have $p_1(u,v) = p_{u_1}(u)p_{v_1}(v)$, and thus

$$\sum_{u\in U_{12}, v\in V_1\cup V_{12}} p(u,v)\log\frac{p(u,v)}{p_u(u)p_v(v)} = 0.$$

Similarly, we have

$$\sum_{u\in U_{12}, v\in V_2} p(u,v)\log\frac{p(u,v)}{p_u(u)p_v(v)} = 0,$$

and thus

$$\sum_{u\in U_{12}, v\in V} p(u,v)\log\frac{p(u,v)}{p_u(u)p_v(v)} = 0,$$

and similarly,

$$\sum_{u\in V_{12}, u\in U} p(u,v)\log\frac{p(u,v)}{p_u(u)p_v(v)} = 0.$$

Thus, we only need to calculate the first two terms:

$$
\begin{aligned}
I(u,v) =& \sum_{i=1}^{2} \sum_{u\in U_i\setminus U_{12}, v\in V_i\setminus V_{12}} \frac{p_i(u,v)}{2} \log \frac{2p_i(u,v)}{p_{u_i}(u)p_i(v)} \\
=& \sum_{u\in U_1\setminus U_{12}, v\in V_1\setminus V_{12}} \frac{p_1(u,v)}{2} + \sum_{u\in U_2\setminus U_{12}, v\in V_2\setminus V_{12}} \frac{p_2(u,v)}{2} \\
=& \frac{1}{2}\left[\sum_{u\in U_1\setminus U_{12}} p_{u_1}(u) \sum_{v\in V_1\setminus V_{12}} p_{v_1}(v) + \sum_{u\in U_2\setminus U_{12}} p_{u_2}(u) \sum_{V\in V_2\setminus V_{12}} p_{v_2}(v) \right] \\
\leq& \frac{1}{4}\left[\sum_{u\in U_1\setminus U_{12}} p_{u_1}(u) + \sum_{v\in V_1\setminus V_{12}} p_{v_1}(v) + \sum_{u\in U_2\setminus U_{12}} p_{u_2}(u) + \sum_{V\in V_2\setminus V_{12}} p_{v_2}(v) \right] \\
=& \frac{1}{4}\left(4 - C_{\text{interleave}}\right).
\end{aligned}
$$

Put together, we have

$$
\begin{aligned}
\overline{I}(u,v) \leq& \frac{4 - C_{\text{interleave}}}{C_{\text{diversity}} + 4 - C_{\text{interleave}}} \\
=& 1 - \frac{C_{\text{diversity}}}{C_{\text{diversity}} + (4 - C_{\text{interleave}})},
\end{aligned}
$$

which completes the proof. □

J Linear Model Analysis for the Impact of Disparity Between Sub-datasets on Shortcut Learning

We consider a simple linear model defined as $\pi_\theta(x) = \pi_\theta([u,v]) = \omega^T[u,v] + b = \omega_1^T u + \omega_2^T v + b$, where the factor generation model g is assumed to be the identity map. We further assume that the sum of prediction errors is zero, i.e., $\mathbb{E}\left[y - \pi_\theta(x)\right] = 0$. This condition can be satisfied by adjusting the bias term b to $\mathbb{E}\left[y - \omega^T[u,v]\right]$. Our focus is on the gradient descent optimization of the parameter ω using the L_2 loss function. The gradients with respect to ω_1 and ω_2 are given by:

$$
[g_{\omega_1}, g_{\omega_2}] = \left[-2\mathbb{E}(y - \omega_1^T u - \omega_2^T v - b)u, -2\mathbb{E}(y - \omega_1^T u - \omega_2^T v - b)v\right].
$$

The magnitudes of these gradients determine the relative importance of the factors u and v in the model's decision-making process. For simplicity, we assume $\mathbb{E}u = \mathbb{E}v = 0$, which does not affect the gradients (by setting $u \leftarrow u - \mathbb{E}u$). Assuming an initial weight of zero, the initial gradient is:

$$
[g_{\omega_1}, g_{\omega_2}] = \left[-2\mathbb{E}((y - \mathbb{E}y)(u - \mathbb{E}u)), -2\mathbb{E}((y - \mathbb{E}y)(v - \mathbb{E}v))\right].
$$

This expression reveals that the gradients measure the correlations between the factors u, v and the target variable y. Importantly, these correlations are strongly influenced by the scale of $u - \mathbb{E}u$ and $v - \mathbb{E}v$. The distance between the factors of sub-datasets significantly affects these scales. Consider a scenario where the distance $d(U_1, U_2)$ is increased to $t \cdot d(U_1, U_2)$ without altering the content of u_1 and u_2. In this case, $\mathbb{E}((y - \mathbb{E}y)(u - \mathbb{E}u))$ will approximately increase to $t \cdot \mathbb{E}((y - \mathbb{E}y)(u - \mathbb{E}u))$, as the increased distance increases the scale and variance of the random variable u by the same extent. Thus, the distances between factors of sub-datasets play a crucial role in determining whether shortcut learning occurs. If spurious correlations exist and the sub-dataset distance of task-irrelevant factors $d(V_1, V_2)$ is significantly greater than that of task-relevant factors $d(U_1, U_2)$, the model is more likely to learn shortcuts.

K Vision and Multimodal Datasets

We list the vision and multimodal datasets we use in Section 2:

ImageNet-1K [91]: ImageNet is a large-scale visual database designed for use in visual tasks. It contains over 14 million images that have been hand-annotated to indicate what objects are picture, and ImageNet-1K is a subset that contains more than 1M images with one thousand classes. It has been widely used for training large-scale vision models, including recent self-supervised models that have been used as the visual encoder for vision-language models.

Open Images [92]: Open Images is a large-scale dataset for object detection, segmentation, and visual relationship detection, containing over 9 million images annotated with image-level labels, object bounding boxes, and visual relationships. It provides a comprehensive resource for developing and benchmarking models in various computer vision tasks, with a focus on real-world image diversity and complexity. We use the sixth version of the dataset.

COCO [93]: The Common Objects in Context (COCO) dataset is a large-scale object detection, segmentation, and captioning dataset. It contains over 330,000 images, with more than 200,000 labeled images and 1.5 million object instances. It has often been used as part of the instruction tuning dataset for vision-language models [19].

ADE20K [94]: ADE20K is a dataset for semantic segmentation and scene parsing, containing over 20,000 images covering a wide range of scenes and object categories. Each image is densely annotated with objects and stuff categories.

iNaturalist [95]: The iNaturalist dataset is a large-scale species classification dataset, derived from the iNaturalist community, which is a citizen science project and online social network of naturalists. It contains millions of images spanning thousands of species.

Flickr30k [96]: Flickr30k is a dataset for multimodal research, consisting of 31,000 images collected from Flickr. Each image is paired with five different captions.

GQA [97]: The GQA (Graph Question Answering) dataset is designed for visual question answering tasks, featuring 22 million questions about 140,000 images. The dataset has been widely used for evaluation of vision-language models, and it has also been used as the tuning dataset of some vision-language models [98].

Visual Genome [99]: Visual Genome is a dataset that connects structured image data with language, containing over 100,000 images with region descriptions, object annotations, attributes, and relationships. It serves as a comprehensive resource for tasks involving scene understanding, object detection, and relationship modeling, facilitating research in bridging vision and language.

LAION-400M [100]: LAION-400M is a large-scale dataset consisting of 400 million image-text pairs, collected from publicly available Common Crawl data. It is designed to support research in large-scale multimodal learning, providing a diverse and extensive resource for training vision-language models.

CC3M [101]: The Conceptual Captions 3M (CC3M) dataset is a large-scale image captioning dataset containing approximately 3.3 million images sourced from the web. Each image is paired with a caption that describes the visual content, offering a valuable resource for training and evaluating models in vision-language tasks.

As there may be overlaps between these datasets, we filter the duplicate data before conducting the analysis in Section 2.