StructPolicy: Structure-Guided Imitation Learning Robust to Visual Domain Shifts

Zehao Du*, Jiude Wei*, Cewu Lu, and Jianhua Sun†

Shanghai Jiao Tong University, Shanghai 200240, China
duzehao_2023jd, wjd_kznwl, lucewu, gothic@sjtu.edu.cn
Homepage: https://structpolicy.github.io/StructPolicy-Homepage/

Abstract. Imitation Learning (IL) offers an effective approach for robot manipulation by learning a mapping from visual inputs to actions. However, this paradigm suffers from poor robustness under visual domain shifts (*e.g.*, lighting conditions, camera viewpoints, *etc.*), often failing to perform manipulation. Our key insight in tackling this problem is to construct a **Structure Map** encoding the fine-grained object structures that remain invariant across visual domains and vital for manipulation. Based on this insight, we propose **StructPolicy**, a method elaborately designed to incorporate the domain-invariant Structure Map into the IL policy through two modules: StructCon and StructEncoder. StructCon is an automated Structure Map construction module that leverages structure primitives to enable flexible transformation and composition to form a wide range of objects. StructEncoder is a hierarchical network that efficiently captures structural relationships and affordance from Structure Map. This provides the IL policy with both domain-invariant and structurally-aware features, guiding it toward robust manipulation under visual domain shifts. We extensively evaluate StructPolicy across 49 manipulation tasks in multiple benchmarks and diverse real-world tasks under various visual changes. The results demonstrate consistent and significant performance improvements across all tasks, validating that StructPolicy enhances the effectiveness and robustness against visual domain shifts of the IL policy, improving manipulation accuracy.

Keywords: Robot Manipulation · Imitation Learning

1 Introduction

Imitation Learning (IL) has emerged as a widely adopted paradigm in robot manipulation due to its efficiency, ease of training, and strong performance across many tasks [26, 31, 50, 52, 77]. However, existing IL methods often fail to generalize beyond the visual conditions of their training demonstrations, exhibiting significant performance degradation under visual domain shifts such as background or viewpoint changes [73, 81]. As shown in Fig. 1-(a) and (b), in a simple

* Equal contribution.
† Corresponding author.

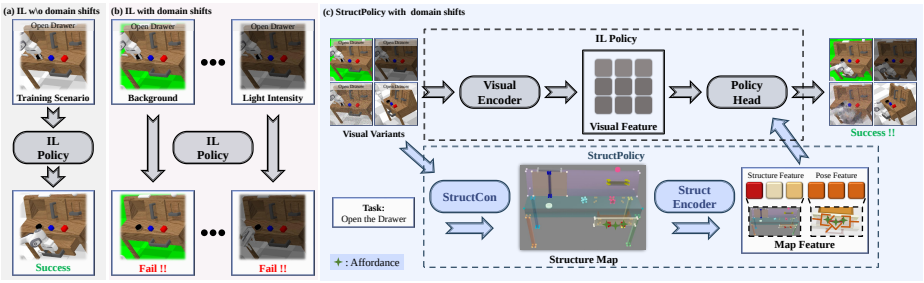

Fig. 1: *(a), (b).* The IL policy fails to complete manipulation tasks (*e.g.*, open the drawer) under visual domain shifts (*e.g.*, background color changes). *(c).* StructPolicy first uses StructCon to construct a Structure Map that encodes fine-grained object structures invariant to visual domain shifts. It then leverages StructEncoder to extract domain-invariant map features that indicate how objects are structured and where they can be manipulated. By providing these features to the IL policy, StructPolicy equips IL policy with a comprehensive understanding of objects to enhance the robustness against visual domain shifts, thereby improving manipulation accuracy.

task to open the drawer, an IL policy that performs successfully in the training scenario can collapse when it encounters visual domain shifts, such as changes in background color and light intensity. These significantly limit the robustness of IL in robotic manipulation scenarios.

Recent works have attempted to tackle this problem by constructing object-centric representations (OCRs), which segment observations, such as RGB images, into distinct entities known as objects [14,43,45,54,67,81,82]. While OCRs effectively filter out irrelevant background information, these representations are *not invariant* to visual domain variations (*e.g.*, illumination, camera viewpoint), which compromises their robustness. In contrast, we propose a novel descriptor that exclusively encodes object structure (*i.e.*, spatial layout and topology). This descriptor remains *invariant* across domain shifts while capturing the features most important for manipulation, guiding IL policies to focus on these invariant structural features. This idea originates from Gibson's ecological theory of perception [21], which posits that affordances grounded in object structure are the primary action-relevant properties perceived by organisms. In this view, manipulation requires a deep understanding of object structure, as it directly determines the possible interactions an agent or human can perform.

To this end, we propose Structure Map, a structure descriptor that models fine-grained object structure as a composition of structure primitives (*e.g.*, cuboids, cylinders). Specifically, the topology of the object structure is modeled by leveraging physical regularities to constrain the sizes and poses of the primitives. Meanwhile, the spatial layout is modeled by the composition and transformation of primitives in 3D space, allowing flexible arrangement while respecting the topology. Together, these perspectives enable Structure Map to model fine-grained object structure beyond coarse spatial or semantic abstractions.

To incorporate Structure Map into IL policy, we elaborately design StructPolicy, a plug-and-play module composed of StructCon and StructEncoder as shown in Fig. 1-(c). *StructCon* leverages structure primitives to enable flexible transformation and composition to form a wide range of objects. It first employs a retrieval-augmented generation (RAG) architecture, where a Vision-Language Model (VLM) takes the task description and observation as input to generate object categories for task-relevant objects, restricted by the ConceptFactory database [59], which stores mathematical rules describing how geometric components are composed to form objects at the category level. This ensures alignment between the VLM's output and categories in ConceptFactory [59], enabling reliable retrieval of the corresponding structure primitives. Subsequently, a parameter estimator estimates the geometric and pose parameters required to transform and compose these primitives under physical constraints, thereby constructing a Structure Map and deriving affordances from it. By leveraging the semantic grounding capability of VLMs and the mathematically encoded structural priors from ConceptFactory [59], StructCon is able to construct structurally correct Structure Maps under visual domain shifts. *StructEncoder* processes the fine-grained object structures encoded in the Structure Map and the derived affordance. For object structure, StructEncoder leverages a hierarchical network to first exploit the geometric properties of each structure primitive and then model the spatial layout and topology relationships of these primitives. For affordance, StructEncoder embeds it to indicate where the robot can interact with the object. By providing features that indicate how an object is structured and where it can be manipulated, StructEncoder enables the IL policy to develop a comprehensive understanding of objects, thereby facilitating more robust manipulation against visual domain shifts.

Our contributions can be summarized as follows: (1) We propose a novel perspective for tackling visual domain shifts in IL, grounded in the insight that object structures remain invariant across visual domains. To leverage this insight, we introduce **Structure Map**—a structure descriptor that encodes fine-grained, domain-invariant object structures, including topology and spatial layout, serving as concrete, domain-invariant knowledge essential for robust manipulation. (2) We propose StructPolicy, a method that incorporates domain-invariant Structure Map to tackle visual domain shifts problem. It first constructs a domain-invariant Structure Map by composing structure primitives under physical constraints, and then extracts features that indicate how an object is structured and where it can be manipulated, thereby equipping IL with a comprehensive object understanding to achieve robustness against visual domain shifts. (3) We conduct comprehensive evaluations across 49 manipulation tasks in multiple benchmarks and diverse real-world tasks, demonstrating consistent performance gains. Further ablation studies under four types of visual changes confirm that StructPolicy notably improves robustness of IL policy against visual domain shifts.

2 Related Work

2.1 Imitation Learning for Robot Manipulation

IL offers an efficient paradigm for robot manipulation by mapping observations to actions from demonstrations without explicit rewards. Early IL methods relate to IRL [1, 40], which infers rewards from demonstrations but suffers from poor sample efficiency in real-world settings. Behavior Cloning (BC) [15,50,51,62] mitigates this by directly learning policies via supervised learning, and has shown strong results in long-horizon and multi-task manipulation [2–4,23,33,42,55]. Recently, generative models have further enhanced IL's expressiveness and stability: Diffusion Policy [16], FlowPolicy [75], and IMLE Policy [46] achieve state-of-the-art manipulation performance. Concurrently, perception-oriented methods like Lift3D [24], PolarNet [12], and Perceiver-Actor [56] exploit RGB-D inputs and 3D spatial encoders to boost robustness. However, most IL approaches are sensitive to visual domain shifts. Our StructPolicy tackles this issue by learning a domain-invariant Structure Map that encodes fine-grained object structures, grounding policy learning in structural knowledge, and improving robustness under visual domain shifts.

2.2 Object-Centric Representation for Robot Manipulation

Object-centric representations have been widely adopted to mitigate domain shifts in IL. This method segments images into distinct entities, known as objects, introducing structure and symbolic reasoning into the representation. Early object-centric methods represent scenes with coarse object-level abstractions. Some works use object proposals or bounding boxes to factor visual observations into object entities and relations [17, 38, 57, 66], while others use object poses or 6-DoF pose estimates as compact states for grasping and manipulation [63, 64]. These abstractions mainly describe objects as whole instances, lacking fine-grained internal structures such as articulated parts, functional surfaces, and compositional topology. Other works learn object-centric embeddings through self-supervised learning [6, 9, 18, 20, 22, 27, 29, 32, 44, 53, 58, 71]. Recent vision foundation models for detection, segmentation, and visual-language reasoning [7, 8, 13, 39, 41, 79] have further enabled object centric representations for robot manipulation [14, 43, 45, 54, 67, 81, 82].

Beyond instance-level object abstractions and learned object-centric embeddings, representing objects through abstract structural components has long been studied in vision, where objects are modeled as compositions of geometric primitives or parts for recognition [5, 34]. Recent robot-learning works have also explored abstract or symbolic object representations for planning and manipulation [30]. StructPolicy follows this structural perspective, but uses compositional object structures as a domain-invariant auxiliary representation for IL, rather than relying only on coarse object instances or learned visual embeddings.

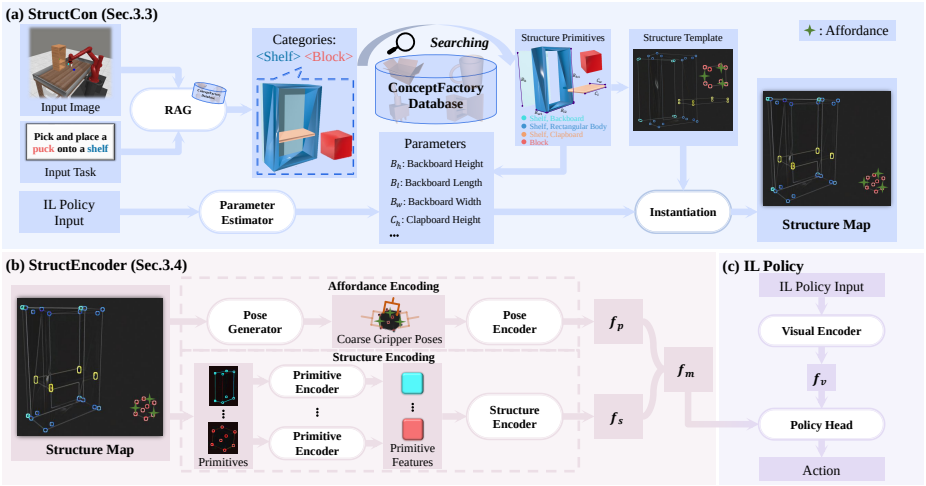

Fig. 2: Overview of StructPolicy. StructPolicy serves as a plug-and-play module that enhances standard imitation learning (IL) policies with StructCon and StructEncoder. Given a task description, an RGB image of the environment, and IL policy input, *StructCon (a)* first constructs a Structure Map that encodes object structures and derived affordance. This Structure Map is then processed by the *StructEncoder (b)* to extract two types of features: the structure feature (f_s) and the coarse gripper pose feature (f_p). The resulting features are fused into map features f_m, which are concatenated with the visual feature f_v from the *IL policy (c)* to predict the final action.

3 Method

3.1 Overview

As illustrated in Fig. 2-(c), most IL policies comprise a visual encoder to extract visual features from observations, and a policy head that maps visual features to actions. StructPolicy is a plug-and-play method that enhances IL by incorporating domain-invariant Structure Map. Specifically, StructPolicy is composed of two modules: StructCon to construct Structure Map and StructEncoder to extract map features from Structure Map.

As shown in Fig. 2-(a), given an RGB image of the environment, a task description and IL policy input, StructCon employs an RAG architecture to generate categories of task-relevant objects and then queries the ConceptFactory [59] database to retrieve the corresponding structure primitives. Subsequently, we leverage mathematical rules of physical constraints and affordance knowledge to form an initial structure template that governs the composition of the retrieved structure primitives. The IL policy input is then passed through parameter estimator to estimate the parameters of the structure template. Finally, these parameters are used to instantiate the structure template, producing the complete Structure Map.

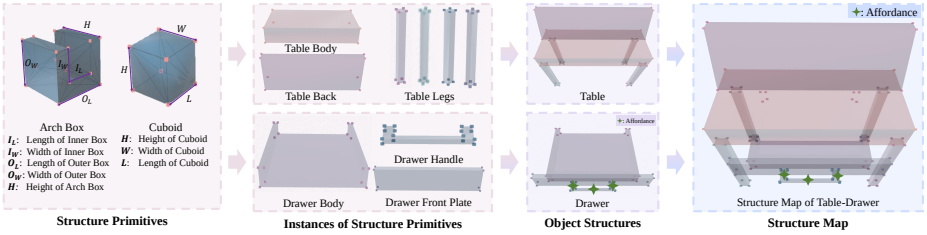

Fig. 3: The Structure Map is organized hierarchically: a set of structure primitives are instantiated into instances of structure primitives, which are grouped to form object structures. The collection of all object structures constitutes the complete Structure Map.

As shown in Fig. 2-(b), StructEncoder extracts map features from the constructed Structure Map using two parallel streams. The affordance encoding stream generates coarse gripper poses—typically located on or pointing toward the affordance regions in the Structure Map—and extracts corresponding pose features. The structure encoding stream first processes intra-primitive geometric properties by primitive encoders to extract primitive features, which are then aggregated to model inter-primitive spatial layout and topology by a structure encoder, yielding the final structure feature. Finally, the poses and structure features are concatenated to form unified map features, which are further concatenated with visual feature as input to the policy head. This design allows the policy to capture structural and affordance knowledge of objects, providing a comprehensive understanding of objects to IL, thereby enhancing IL's effectiveness and robustness against visual domain shifts.

3.2 Structure Map

The Structure Map is formulated in three levels as shown in Fig. 3: structure primitives, object structure, and the final Structure Map.

Structure Primitives. We begin with a library of structure primitives (*e.g.*, *Cuboid*, *Sphere*) adapted from geometry templates proposed by Sun *et al.* [59]. To construct structure primitives, each shape of geometry template is sampled into labeled 3D points using two structured sampling strategies. For polyhedral geometries (*e.g.*, *Cuboid*), we uniformly triangulate the surface and take triangle vertices as sampled points. For curved primitives (*e.g.*, *Cylinder*), we sample points on planar arc sections with a sampling resolution N. These sampled point sets serve as reusable templates called structure primitives for building more complex structures.

Object Structure. Formally, the object structure $\mathcal{S}_\mathbf{o}$ is defined as the union of the corresponding structure primitives to model spatial layout:

$$\mathcal{S}_\mathbf{o} = \bigcup_{m=1}^{N_\mathbf{o}} \{(\mathbf{p}_i, \ell_i)\}_{i=1}^{N_m}, \tag{1}$$

where $\mathbf{p}_i \in \mathbb{R}^3$ is a sampled point, ℓ_i specifies the type of the associated structure primitive, N_m is the number of points sampled from the m-th structure primitive, and $N_{\mathbf{o}}$ is the total number of primitives. The modeling of topology is achieved by further imposing physical constraints on each primitive according to ℓ_i.

Affordance. The affordance $\mathcal{A}_{\mathbf{o}}$ is defined as the set of triangular surface patches that correspond to actionable regions of the object: $\mathcal{A}_{\mathbf{o}} = \bigcup_{i=1}^{K_{\mathbf{o}}} \mathcal{F}_i$, where $\mathcal{F}_i = \{\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, \mathbf{v}_{i,3} \mid \mathbf{v}_{i,j} \in \mathbb{R}^3\}$. Each $\mathcal{F}_i$ represents a triangular face on the object surface that affords a specific manipulation behavior, and $K_{\mathbf{o}}$ is the total number of such affordance-relevant faces. The affordance $\mathcal{A}_{\mathbf{o}}$ is computed from the affordance knowledge described in Sun *et al*. [59].

Structure Map. Given a task environment with K task-relevant objects, Structure Map is defined as the collection of object structures and their affordance:

$$\mathcal{M} = \{(\mathcal{S}_{\mathbf{o}_k}, \mathcal{A}_{\mathbf{o}_k}, k)\}_{k=1}^{K}. \tag{2}$$

Thus, the Structure Map encodes both object structures and derived affordance in a unified representation.

3.3 StructCon

StructCon is responsible for automatically constructing Structure Map, taking as input the task description, the RGB image of the environment, and the full input provided to IL. StructCon begins by employing RAG to generate categories of task-relevant objects in the scene. Specifically, the RGB image of the scene and the task description are first processed by a VLM to generate a textual summary of task-relevant objects. This summary is encoded into a dense query embedding using BGE-M3 [11]. The resulting embedding is then used to retrieve the most semantically relevant object categories from the pre-indexed ConceptFactory database [59] via dense vector similarity search. These retrieved categories are subsequently used to query the ConceptFactory database [59], which stores category-template key-value pairs, retrieving the corresponding structure primitives, which are then composed into an initial structure template governed by physical constraints according to templates in ConceptFactory Suite [59]. Finally, a parameter estimator, where IL policy input is first encoded into a visual feature through the visual encoder from IL policy and then fed into an MLP head, estimates the parameters of the structure template. These parameters are then used to instantiate the Structure Map that encodes object structures and derived affordance. Note that the ConceptFactory database [59] is a large collection of 4380 objects acquired from 39 categories, enabling StructCon to be adaptable to diverse manipulation scenarios.

3.4 StructEncoder

To incorporate Structure Map into the downstream policy head, we introduce **StructEncoder**, which extracts map features from the Structure Map. StructEncoder consists of two feature extraction streams: one focuses on affordance,

and the other hierarchically encodes object structure.

Affordance Encoding Stream. The affordance encoding stream operates on the *coarse gripper poses*, which represent a set of candidate gripper poses on manipulable objects derived from the Structure Map. These coarse poses provide affordance guidance for manipulation. Formally, the set of coarse poses is defined as

$$\mathcal{P} = \{(x_i, y_i, z_i, d_i^x, d_i^y, d_i^z) \mid \|(d_i^x, d_i^y, d_i^z)\|_2 = 1\}_{i=1}^{K_{\text{pose}}} \tag{3}$$

where (x_i, y_i, z_i) denotes the position of the i-th coarse pose and (d_i^x, d_i^y, d_i^z) is a unit direction vector indicating the approach direction of the gripper and pointing toward the corresponding affordance region.

A pose generator takes the affordance in Structure Map as input and produces the coarse poses $\mathcal{P}$. For each affordance in the Structure Map, its geometric center is used as the initial coarse pose position (x, y, z), and the inverse of its surface normal is assigned as the initial orientation (d^x, d^y, d^z). After that, an MLP predicts a small position offset $\Delta(x, y, z)$ for each coarse pose. The final position is computed as $(x, y, z) \leftarrow (x, y, z) + \Delta(x, y, z)$, while the orientation remains unchanged, enabling more precise and task-relevant coarse pose placement.

Finally, the coarse gripper poses are processed by an MLP-based *pose encoder* for the pose feature f_p, capturing affordance for downstream policy learning.

Structure Encoding Stream. The structure encoding stream encodes the structural information from Structure Map. A hierarchical encoding strategy is adopted to capture both intra-primitive geometric properties and inter-primitive topology and spatial layout. Specifically, the Structure Map is first partitioned into groups based on primitive labels, and each group is independently processed by a *primitive encoder*, leveraging MLPs followed by attention pooling to produce local primitive features that encode geometric properties specific to each primitive type. The resulting primitive features are subsequently aggregated and passed into *structure encoder*, leveraging transformers to capture inter-primitive structural knowledge including topology and spatial relationships across objects. The final output of this stream is a global structure feature f_s that integrates both local geometric patterns and global structural relationships.

Feature Fusion. The pose feature f_p and the global structure feature f_s are concatenated and then passed through an MLP to form the final map feature f_m, which is fed into the downstream IL policy head.

3.5 Implementation Details

All MLPs within StructPolicy are implemented using three linear layers with ReLU activation. The Transformer components are adopted from the architecture proposed by Zhao *et al.* [76]. StructPolicy is integrated and trained jointly with the IL backbone, following the original IL training settings and requiring no additional loss terms.

Table 1: Comparison of StructPolicy and baselines on CALVIN [37] and MetaWorld [35, 69]. Δ denotes the improvement of our StructPolicy over the baseline.

Config	CALVIN [37] D → D (Avg. Len.)					MetaWorld [35, 69] (S.R. %)		
	TaKSIE [25]	HULC++ [36]	MDT [48]	MDT-V [48]	FLOWER [49]	DP [16]	DP3 [74]	Lift3D [24]
Baseline	3.18	3.30	3.59	3.72	4.35	65.2	66.7	83.9
+ StructPolicy	**3.34**	**3.45**	**3.70**	**3.86**	**4.42**	**77.5**	**79.2**	**90.5**
Improvement (Δ)	+0.16	+0.15	+0.11	+0.14	+0.07	+12.3	+12.5	+6.6

Config	CALVIN ABCD → D (Avg. Len.)					
	DeeR [72]	GR-1 [68]	MoDE [47]	MDT [48]	MDT-V [48]	FLOWER [49]
Baseline	4.13	4.21	4.39	4.33	4.51	4.67
+ StructPolicy	**4.22**	**4.28**	**4.45**	**4.49**	**4.60**	**4.70**
Improvement (Δ)	+0.09	+0.07	+0.06	+0.16	+0.09	+0.03

4 Experiments

In Sec. 4.1, we validate the effectiveness of StructPolicy through simulated experiments. The robustness of StructPolicy against visual domain shifts is evaluated in Sec. 4.2. We further quantify the contribution of each component of StructCon in Sec. 4.3. We validate the effectiveness of the Structure Map representation and StructEncoder through ablation studies in Sec. 4.4. The inference efficiency of StructPolicy is evaluated in Sec. 4.5. We also deployed StructPolicy in real-world settings in Sec. 4.6.

4.1 Effectiveness of StructPolicy

Experiment Settings. We benchmark using two standard manipulation environments MetaWorld [35, 69] and CALVIN [37]. (1) **MetaWorld** [35, 69] is a MuJoCo-based [61] benchmark with 50 manipulation tasks across four difficulty levels. Following Jia *et al.* [24], we select 15 representative tasks spanning easy (*button-press, drawer-open, reach, handle-pull, peg-unplug-side, lever-pull, dial-turn*), medium (*hammer, sweep-into, bin-picking, push-wall, box-close*), and hard/very hard (*assembly, hand-insert, shelf-place*). Training uses 30 demonstrations per task, 100 epochs with rollouts every 10, and reports the maximum average success rate across tasks, evaluated from two camera viewpoints and averaged. (2) **CALVIN** [37] features a multifunctional table and three blocks varying in sizes across four environments (A, B, C, D). Two standard settings are considered: $ABCD \rightarrow D$ (trained on 24 hours of uncurated play data covering 34 tasks across all four environments, tested on D) and $D \rightarrow D$ (trained and evaluated on D with a smaller 6-hour dataset). Evaluation follows the long-horizon benchmark: 1000 natural language instruction chains, each requiring continuous completion of 5 tasks. Robots receive a reward of 1 per successful instruction (max score of 5 per rollout).

Results. MetaWorld [35,69] comprises a suite of tasks that demand fine-grained interactions with articulated objects (*e.g.*, *lever-pull*) or precise placement of objects (*e.g.*, *assembly*). Successfully performing these tasks requires the policy to

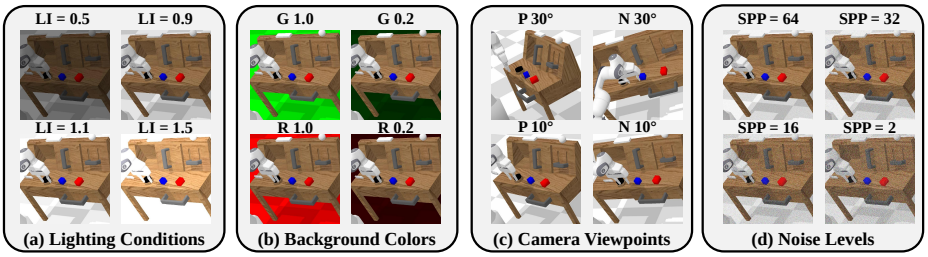

Fig. 4: Examples of visual changes. *(a)* LI indicates light intensity. *(b)* G and R mean green or red and the number represents the value of the green or red channel. *(c)* P denotes positive and N denotes negative. *(d)* SPP means samples per pixel in the ray tracing.

understand the fine-grained structure of objects to determine where to interact with them. StructPolicy leverages such object structural knowledge, making MetaWorld [35, 69] an ideal benchmark for demonstrating its advantages. As shown in Tab. 1, StructPolicy consistently outperforms all baselines. Notably, it is the first policy, to the best of our knowledge, to exceed the 90% success rate threshold. The CALVIN [37] benchmark is built around a topology-rich tabletop environment that features articulated objects such as drawers and sliding panels, requiring the robot to execute a sequence of actions (*e.g.*, open the drawer, open the slide) in response to instructions. Successfully completing these tasks hinges on the policy's ability to understand object structure—which parts are movable, how they are connected, and where to interact with them. This setting thus provides an ideal testbed for StructPolicy, which leverages object structures and affordance. As shown in Tab. 1, StructPolicy demonstrates clear advantages, significantly outperforming prior methods under both evaluation protocols, achieving an average instruction chain length of 4.70 on the ABCD $\rightarrow$ D challenge (surpassing FLOWER [49]'s 4.67) and 4.42 on the D $\rightarrow$ D setting (beating FLOWER [49]'s 4.35). These results confirm that StructPolicy effectively guides IL to master complex, structured manipulation tasks.

4.2 Robustness to Visual Domain Shifts

We evaluate the robustness of StructPolicy against visual domain shifts in the CALVIN benchmark using MDT-V as the backbone [37, 48]. As shown in Fig. 4, we evaluate policy under four types of visual changes following Zhu *et al.* [80]: *1) Lighting* with intensity levels 0.5–1.5 relative to the default. *2) Background color* with red or green channel values varied across 0.2, 0.6, and 1.0. *3) Camera viewpoint* with horizontal shifts of 10°, 20°, and 30° and *4) visual noise* by switching the rendering mode to ray tracing, disabling the denoiser, and varying the number of ray tracing samples per pixel to 2, 16, 32, and 64.

As shown in Fig. 5, StructPolicy consistently outperforms MDT-V [48] across all settings. This robustness to visual domain shifts stems from Structure Map

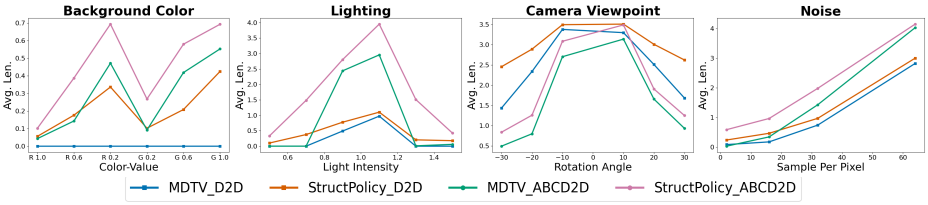

Fig. 5: Quantitative results of policy performance under visual domain shifts in the CALVIN [37] environment. StructPolicy consistently outperforms MDT-V [48] across all types of domain shifts, including lighting, background color, camera viewpoint, and noise.

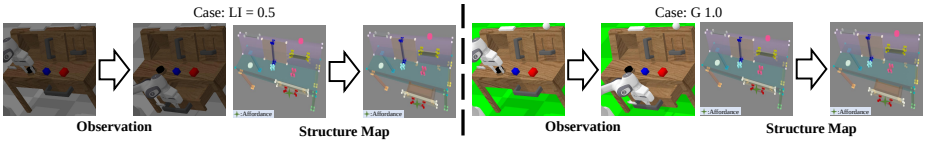

Fig. 6: Case study of StructPolicy under visual domain shifts with task "open the drawer". *LI* refers to Light Intensity, and *G 1.0* denotes a background color with green channel value of 1.0.

that encodes domain-invariant object structures. This invariance is illustrated in Fig. 6 for the *open the drawer* task: despite changes in observation—such as dark lighting or a green background—Structure Map constructed by StructPolicy remains consistent, preserving the drawer's geometric articulation, pull affordance, and relative pose with respect to the table. By grounding IL in domain-invariant structural knowledge, StructPolicy equips the IL policy with a comprehensive understanding of how the object is structured and where it can be manipulated, enhancing its robustness against visual domain shifts.

We measure the accuracy of constructed Structure Maps under visual domain shifts in the CALVIN [37] benchmark. Specifically, we use SAM [28] to extract masks of the table and blocks, which are converted into ground-truth point clouds in the world coordinate frame. Predicted point clouds are created by uniformly sampling points from the constructed Structure Maps. We quantify geometric deviation using Chamfer Distance (CD) [19, 70], Earth Mover's Distance (EMD) [70], and F-scores [60] between the predicted and ground-truth point clouds. As shown in Tab. 2, StructCon maintains accuracy: compared to the no-change case, the average CD ($\times 10^{-3}$) rises from 8.66 to an average of 12.89 across the four shifts, and the average EMD ($\times 10^{-2}$) slightly increases from 4.56 to an average of 7.07. Crucially, the F-Score remains high, decreasing only modestly from 61.6% to an average of 51.8%. These results strongly confirm that StructCon consistently constructs accurate, structure-preserving maps under visual domain shifts, primarily due to: (1) Robust cross-modal VLM grounding for accurate task-relevant object identification, and (2) strong struc-

Table 2: Evaluation of structure map construction accuracy under visual domain shifts in the CALVIN [37] benchmark. CD, EMD and F-Score are used to measure the geometric deviation between the constructed structure map and the ground-truth point cloud.

Visual Variant	**CD** ($\times 10^{-3}$)↓	**EMD** ($\times 10^{-2}$)↓	**F-Score (%)**↑
No Change	**8.66**	**4.56**	**61.6**
Lighting (avg)	12.23	6.79	53.7
Background Color (avg)	12.90	7.01	51.3
Camera Viewpoint (avg)	13.55	7.56	50.2
Noise (avg)	12.89	6.93	52.1

Table 3: Error Breakdown of StructCon components in the MetaWorld [35,69] benchmark. "✓" indicates ground-truth information was provided for that component, while a dash (–) denotes that the output from the corresponding estimation module was used.

Line	Structure Template	Parameters	Mean S.R.	Δ (Rel.)
0	–	–	90.5	-
1	✓	–	91.0	+0.55%
2	✓	✓	98.4	+8.73%

tural priors provided by describing object shapes with geometric primitives via ConceptFactory [59].

4.3 System Error Breakdown

To quantify the contribution of each component of StructCon to the overall manipulation accuracy, we perform a stage-wise error breakdown, where ground-truth is progressively introduced for different components. As shown in Tab. 3, providing the ground-truth for the structure template only improves the success rate by 0.5%, suggesting that the RAG is highly accurate in identifying the object categories and contributes minimal error to the system. In contrast, providing the ground-truth for the parameters causes the success rate to jump from 91.0% to 98.4%. This significant 7.4% gain clearly indicates that the parameter estimator is the primary performance bottleneck of StructCon.

4.4 Ablation Studies

Effectiveness of StructEncoder & Structure Map. Table 4 evaluates the impact of StructEncoder and the Structure Map on MetaWorld [35, 69]. We compare StructEncoder against three baselines—PointNet, PointTransformer, and DenseFusion—using both the structured sampling described in Sec. 3.2 and uniform point sampling [10,65,76]. Structured sampling consistently outperforms uniform sampling across all encoders. StructEncoder achieves the highest success rate (90.5%) when paired with structured sampling, confirming the hierarchical

Table 4: Ablation study on Map Encoder and Representation.

	Map Encoder				Map Representation		
	SE	PT	DF	PN	Structure	Uniform	**Mean**
Ex1	-	-	-	✓	-	✓	**79.8**
Ex2	-	-	-	✓	✓	-	**80.3**
Ex3	-	-	✓	-	-	✓	**83.2**
Ex4	-	-	✓	-	✓	-	**84.0**
Ex5	-	✓	-	-	-	✓	**79.6**
Ex6	-	✓	-	-	✓	-	**80.8**
Ex7	✓	-	-	-	-	✓	**80.5**
Ex8	✓	-	-	-	✓	-	**90.5**

Table 5: Ablation study on the design of StructEncoder.

	Structure			Affordance		Fusion			
	Ours	PT	PN	Ours	MLP	Concat	Ours	Attn	**Mean**
Ex1	-	-	-	✓	-	-	-	-	**74.6**
Ex2	✓	-	-	-	-	-	-	-	**85.2**
Ex3	-	-	✓	✓	-	-	✓	-	**81.2**
Ex4	-	✓	-	✓	-	-	✓	-	**84.5**
Ex5	✓	-	-	-	✓	-	✓	-	**86.3**
Ex6	✓	-	-	✓	-	✓	-	-	**87.1**
Ex7	✓	-	-	✓	-	-	-	✓	**88.4**
Ex8	✓	-	-	✓	-	-	✓	-	**90.5**

Note: SE = StructEncoder (ours), PT = PointTransformer [76], DF = DenseFusion [65], PN = PointNet [10]. Structure = structured sampling (ours), Uniform = uniform surface sampling. Darker blue indicates better performance.

Table 6: Ablation study on the contribution of Structure Map compared with simpler object-centric representations on MetaWorld [35,69]. We report mean success rates (%) over 15 tasks.

Backbone	Base	+Mask	+Keypoints	+Pose	+BBox	+Structure Map
DP [16]	65.2	69.6	71.1	67.8	70.2	**77.5**
Improvement (Δ)	-	+4.4	+5.9	+2.6	+5.0	+12.3
DP3 [74]	66.7	70.2	70.8	73.6	69.4	**79.2**
Improvement (Δ)	-	+3.5	+4.1	+6.9	+2.7	+12.5

network's ability to capture structural features effectively.

Contribution of Structure Map. Table 6 compares Structure Map with simpler object-centric representations on MetaWorld [35, 69], including object masks, keypoints, poses, and bounding boxes. These representations provide task-relevant object cues but do not explicitly model the compositional structure, topology, and affordance encoded by Structure Map. Results show that all auxiliary representations improve the base policies, indicating that object-centric information is helpful for manipulation. However, Structure Map achieves the largest gains on both backbones, improving DP [16] by 12.3% and DP3 [74] by 12.5%. These results show that the performance gain of StructPolicy comes not merely from adding object-level cues, but from the fine-grained structural knowledge provided by Structure Map.

Design of StructEncoder. Table 5 analyzes the specific encoding and fusion components within StructEncoder. Results indicate: *1) Encoding method:* Our specific design (Ex8) outperforms alternative structure encoders (PointNet [10], PointTransformer [76]) and affordance encoders (simple MLP). *2) Component necessity:* Combining structure and affordance (Ex8) yields significant gains over using either feature in isolation (Ex1, Ex2). *3) Feature fusion:* Our MLP-based fusion (Ex8) integrates multi-modal features more effectively than concatenation (Ex6) or attention mechanisms (Ex7).

Table 7: Inference Speed of Lift3D [24] and StructPolicy over 15 tasks in Meta-World [35, 69].

Method	Lift3D (CLIP) [24]	StructPolicy
Average Inference Time (ms) ↓	67.21	70.20 (+2.99)

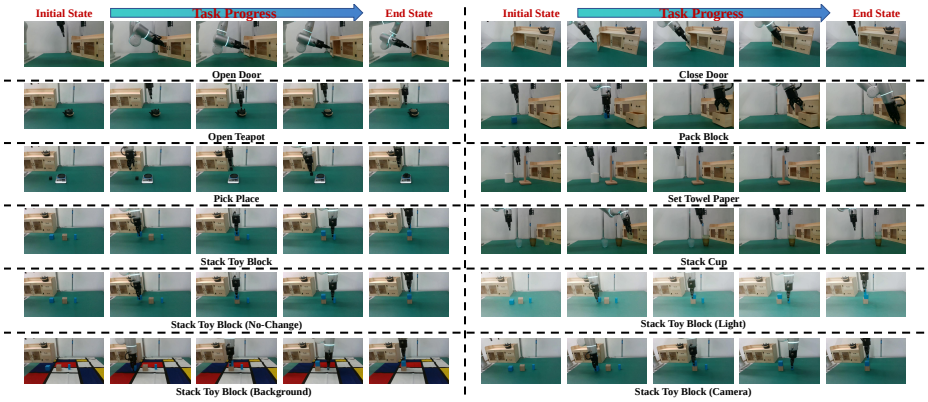

Fig. 7: Visualization of real-world experiments across eight tasks, with an example of four visual domains for the *Stack Toy Block* task.

4.5 Inference Efficiency of StructPolicy

To assess the inference cost of our approach, we report inference speeds averaged over 15 MetaWorld [35,69] tasks, as summarized in Tab. 7. StructPolicy exhibits a negligible overhead of only 2.99 milliseconds, compared to the baseline Lift3D (CLIP) model [24], demonstrating that the introduced structured representation incurs minimal runtime cost while enabling improved performance.

4.6 Real-World Experiments

We deploy StructPolicy on a Flexiv Rizon 4 robotic arm with a 7-DoF configuration and a Flexiv-GN01 gripper, operating at 30 Hz. RGB-D observations are captured using an Intel RealSense L515, and point clouds are sampled to 1024 points using Open3D [78]. We adopt four policies as backbones: DP [16], DP3 [74], MDT-V [48], and Lift3D [24].

As shown in Fig. 7, we evaluate StructPolicy on eight real-world manipulation tasks covering diverse skills—*Open Door, Close Door, Open Teapot, Pack Block, Pick Place, Set Towel Paper, Stack Toy Block and Stack Cup*—spanning pick-and-place, pushing, stacking, articulated-object control, and precise placement. Each method is tested over 20 trials per task. To evaluate robustness against visual domain shifts, we further evaluate StructPolicy against Lift3D [24] across variations in light intensity, background color, and camera viewpoint, as shown in Fig. 7. All eight tasks are performed under these settings.

Table 8: Real-world quantitative results across different policy backbones. We report success rates (%).

Method	DP [16]	DP3 [74]	MDT-V [48]	Lift3D [24]
Base	31.9	48.1	58.1	67.5
+ **Ours**	**50.6**	**66.9**	**74.4**	**83.3**
Improv.	*+18.7*	*+18.8*	*+16.3*	*+15.8*

Table 9: Robustness analysis under visual domain shifts using the Lift3D [24] backbone. (Success rates %).

Domain	No-Change	Light	Background	Camera
Lift3D [24]	67.5	30.0	28.3	24.2
+ **Ours**	**83.3**	**65.8**	**69.1**	**60.0**
Improv.	*+15.8*	*+35.8*	*+40.8*	*+35.8*

Table 8 shows that StructPolicy consistently outperforms all baselines across different policy backbones, with average success rate improvements ranging from +15.8% to +18.8%. As shown in Tab. 9, this advantage is particularly pronounced under various visual domain shifts: StructPolicy maintains high success rates compared to the original Lift3D [24], demonstrating superior robustness against lighting variations (65.8% vs. 30.0%), background color changes (69.1% vs. 28.3%), and camera viewpoint shifts (60.0% vs. 24.2%). These results validate the advantages of our approach in guiding the policy to learn domain-invariant object structures in real-world scenarios. Notably, while original baselines suffer drastic performance drops under challenging visual perturbations—particularly in camera viewpoint shifts where success rates fall below 40%—StructPolicy retains significant stability, effectively bridging the performance gap and ensuring reliable robustness.

5 Conclusion

In this work, we propose a novel perspective for tackling visual domain shifts in IL, grounded in the insight that object structures remain invariant across visual domains. To leverage this insight, we introduce Structure Map, a descriptor encoding fine-grained, domain-invariant object structures. To incorporate Structure Map into IL, we design StructPolicy, a plug-and-play method composed of StructCon and StructEncoder. StructCon is an automated Structure Map construction module that leverages structure primitives to enable flexible transformation and composition to form diverse objects. StructEncoder is a hierarchical network that efficiently captures structural relationships and affordance from Structure Map. We conduct comprehensive experiments across 49 manipulation tasks in multiple benchmarks and diverse real-world tasks under visual domain shifts, demonstrating that StructPolicy significantly enhances IL's effectiveness and robustness against visual domain shifts.

6 Acknowledgements

This work was supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China, the Shanghai Committee of Science and Technology, China (Grant No. 24511103200), the Shanghai Artificial Intelligence Laboratory, and XPLORER PRIZE grants.

References

1. Abbeel, P., Ng, A.Y.: Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the Twenty-First International Conference on Machine Learning. p. 1 (2004). https://doi.org/10.1145/1015330.1015430
2. Bharadhwaj, H., Dwibedi, D., Gupta, A., Tulsiani, S., Doersch, C., Xiao, T., Shah, D., Xia, F., Sadigh, D., Kirmani, S.: Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. In: Proceedings of The 9th Conference on Robot Learning. vol. 305, pp. 3936–3951 (2025), https://proceedings.mlr.press/v305/bharadhwaj25a.html
3. Bharadhwaj, H., Mottaghi, R., Gupta, A., Tulsiani, S.: Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXVI. pp. 306–324 (2024). https://doi.org/10.1007/978-3-031-73116-7_18
4. Bharadhwaj, H., Vakil, J., Sharma, M., Gupta, A., Tulsiani, S., Kumar, V.: Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 4788–4795 (2024). https://doi.org/10.1109/ICRA57147.2024.10611293
5. Biederman, I.: Recognition-by-components: A theory of human image understanding. Psychological Review **94**(2), 115–147 (1987). https://doi.org/10.1037/0033-295X.94.2.115
6. Bruno Sauvalle, A.d.L.F.: Unsupervised multi-object segmentation using attention and soft-argmax. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3266–3275 (2023). https://doi.org/10.1109/WACV56688.2023.00328
7. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6154–6162 (2018). https://doi.org/10.1109/CVPR.2018.00644
8. Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9630–9640 (2021). https://doi.org/10.1109/ICCV48922.2021.00951
9. Chang, M., Griffiths, T.L., Levine, S.: Object representations as fixed points: training iterative refinement algorithms with implicit differentiation. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (2022)
10. Charles, R.Q., Su, H., Kaichun, M., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 77–85 (2017). https://doi.org/10.1109/CVPR.2017.16
11. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In: Findings of the Association for Computational Linguistics: ACL 2024 (2024). https://doi.org/10.18653/v1/2024.findings-acl.137
12. Chen, S., Pinel, R.G., Schmid, C., Laptev, I.: Polarnet: 3d point clouds for language-guided robotic manipulation. In: Proceedings of The 7th Conference on Robot Learning. vol. 229, pp. 1761–1781 (2023), https://proceedings.mlr.press/v229/chen23b.html

13. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. vol. 119, pp. 1597–1607 (2020), https://proceedings.mlr.press/v119/chen20j.html
14. Chen, Y., Chen, Z., Yin, J., Huo, J., Tian, P., Shi, J., Gao, Y.: Gravmad: Grounded spatial value maps guided action diffusion for generalized 3d manipulation. In: International Conference on Learning Representations. vol. 2025, pp. 102042–102074 (2025), https://proceedings.iclr.cc/paper_files/paper/2025/file/fce176458ff542940fa3ed16e6f9c852-Paper-Conference.pdf
15. Chen, Y., Wang, C., Fei-Fei, L., Liu, K.: Sequential dexterity: Chaining dexterous policies for long-horizon manipulation. In: Proceedings of The 7th Conference on Robot Learning. vol. 229, pp. 3809–3829 (2023), https://proceedings.mlr.press/v229/chen23e.html
16. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research (2024). https://doi.org/10.1177/02783649241273668
17. Devin, C., Abbeel, P., Darrell, T., Levine, S.: Deep object-centric representations for generalizable robot learning. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 7111–7118 (2018). https://doi.org/10.1109/ICRA.2018.8461196
18. Engelcke, M., Kosiorek, A.R., Jones, O.P., Posner, I.: Genesis: Generative scene inference and sampling with object-centric latent representations. In: International Conference on Learning Representations (2019)
19. Fan, H., Su, H., Guibas, L.: A point set generation network for 3d object reconstruction from a single image. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2463–2471 (2017). https://doi.org/10.1109/CVPR.2017.264
20. Fan, K., Bai, Z., Xiao, T., Zietlow, D., Horn, M., Zhao, Z., Simon-Gabriel, C.J., Shou, M.Z., Locatello, F., Schiele, B., Brox, T., Zhang, Z., Fu, Y., He, T.: Unsupervised open-vocabulary object localization in videos. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13701–13709 (2023). https://doi.org/10.1109/ICCV51070.2023.01264
21. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin (1979). https://doi.org/10.4324/9781315740218
22. Greff, K., Kaufman, R.L., Kabra, R., Watters, N., Burgess, C., Zoran, D., Matthey, L., Botvinick, M., Lerchner, A.: Multi-object representation learning with iterative variational inference. In: Proceedings of the 36th International Conference on Machine Learning. vol. 97, pp. 2424–2433 (2019), https://proceedings.mlr.press/v97/greff19a.html
23. Haldar, S., Peng, Z., Pinto, L.: Baku: an efficient transformer for multi-task policy learning. In: Proceedings of the 38th International Conference on Neural Information Processing Systems (2025)
24. Jia, Y., Liu, J., Chen, S., Gu, C., Wang, Z., Luo, L., Li, X., Wang, P., Wang, Z., Zhang, R., Zhang, S.: Lift3d policy: Lifting 2d foundation models for robust 3d robotic manipulation. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17347–17358 (2025). https://doi.org/10.1109/CVPR52734.2025.01617
25. Kang, X., Kuo, Y.L.: Incorporating task progress knowledge for subgoal generation in robotic manipulation through image edits. In: 2025 IEEE/CVF Winter

Conference on Applications of Computer Vision (WACV). pp. 7490–7499 (2025). https://doi.org/10.1109/WACV61041.2025.00728
26. Kim, H., Ohmura, Y., Kuniyoshi, Y.: Transformer-based deep imitation learning for dual-arm robot manipulation. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8965–8972 (2021). https://doi.org/10.1109/IROS51168.2021.9636301
27. Kipf, T., van der Pol, E., Welling, M.: Contrastive learning of structured world models. In: International Conference on Learning Representations (2019)
28. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3992–4003 (2023). https://doi.org/10.1109/ICCV51070.2023.00371
29. Kulkarni, T., Gupta, A., Ionescu, C., Borgeaud, S., Reynolds, M., Zisserman, A., Mnih, V.: Unsupervised learning of object keypoints for perception and control. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems (2019)
30. Liang, Y., Kumar, N., Tang, H., Weller, A., Tenenbaum, J.B., Silver, T., Henriques, J.F., Ellis, K.: Visualpredicator: Learning abstract world models with neuro-symbolic predicates for robot planning. In: International Conference on Learning Representations (2025)
31. Lin, X., So, J., Mahalingam, S., Liu, F., Abbeel, P.: Spawnnet: Learning generalizable visuomotor skills from pre-trained network. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 4781–4787 (2024). https://doi.org/10.1109/ICRA57147.2024.10610356
32. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (2020)
33. Mandlekar, A., Xu, D., Martín-Martín, R., Savarese, S., Fei-Fei, L.: GTI: Learning to Generalize across Long-Horizon Tasks from Human Demonstrations. In: Proceedings of Robotics: Science and Systems (2020). https://doi.org/10.15607/RSS.2020.XVI.061
34. Marr, D., Nishihara, H.K.: Representation and recognition of the spatial organization of three-dimensional shapes. Proceedings of the Royal Society of London. Series B, Biological Sciences **200**(1140), 269–294 (1978), http://www.jstor.org/stable/77391
35. McLean, R., Chatzaroulas, E., McCutcheon, L., Röder, F., Yu, T., He, Z., Zentner, K., Julian, R., Terry, J.K., Woungang, I., Farsad, N., Castro, P.S.: Meta-world+: An improved, standardized, RL benchmark. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
36. Mees, O., Borja-Diaz, J., Burgard, W.: Grounding language with visual affordances over unstructured data. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 11576–11582 (2023). https://doi.org/10.1109/ICRA48891.2023.10160396
37. Mees, O., Hermann, L., Rosete-Beas, E., Burgard, W.: Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robotics and Automation Letters **7**(3), 7327–7334 (2022). https://doi.org/10.1109/LRA.2022.3180108

38. Migimatsu, T., Bohg, J.: Object-centric task and motion planning in dynamic environments. IEEE Robotics and Automation Letters **5**(2), 844–851 (2020). https://doi.org/10.1109/LRA.2020.2965875
39. Nair, S., Rajeswaran, A., Kumar, V., Finn, C., Gupta, A.: R3m: A universal visual representation for robot manipulation. In: Proceedings of The 6th Conference on Robot Learning. vol. 205, pp. 892–909 (2022), https://proceedings.mlr.press/v205/nair23a.html
40. Ng, A.Y., Russell, S.: Algorithms for inverse reinforcement learning. In: Proceedings of the Seventeenth International Conference on Machine Learning. pp. 663–670 (2000)
41. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. (2024)
42. O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., Tung, A., Bewley, A., Herzog, A., Irpan, A., Khazatsky, A., Rai, A., Gupta, A., Wang, A., Singh, A., Garg, A., Kembhavi, A., Xie, A., Brohan, A., Raffin, A., Sharma, A., Yavary, A., Jain, A., Balakrishna, A., Wahid, A., Burgess-Limerick, B., Kim, B., Schölkopf, B., Wulfe, B., Ichter, B., Lu, C., Xu, C., Le, C., Finn, C., Wang, C., Xu, C., Chi, C., Huang, C., Chan, C., Agia, C., Pan, C., Fu, C., Devin, C., Xu, D., Morton, D., Driess, D., Chen, D., Pathak, D., Shah, D., Büchler, D., Jayaraman, D., Kalashnikov, D., Sadigh, D., Johns, E., Foster, E., Liu, F., Ceola, F., Xia, F., Zhao, F., Stulp, F., Zhou, G., Sukhatme, G.S., Salhotra, G., Yan, G., Feng, G., Schiavi, G., Berseth, G., Kahn, G., Wang, G., Su, H., Fang, H.S., Shi, H., Bao, H., Ben Amor, H., Christensen, H.I., Furuta, H., Walke, H., Fang, H., Ha, H., Mordatch, I., Radosavovic, I., Leal, I., Liang, J., Abou-Chakra, J., Kim, J., Drake, J., Peters, J., Schneider, J., Hsu, J., Bohg, J., Bingham, J., Wu, J., Gao, J., Hu, J., Wu, J., Wu, J., Sun, J., Luo, J., Gu, J., Tan, J., Oh, J., Wu, J., Lu, J., Yang, J., Malik, J., Silvério, J., Hejna, J., Booher, J., Tompson, J., Yang, J., Salvador, J., Lim, J.J., Han, J., Wang, K., Rao, K., Pertsch, K., Hausman, K., Go, K., Gopalakrishnan, K., Goldberg, K., Byrne, K., Oslund, K., Kawaharazuka, K., Black, K., Lin, K., Zhang, K., Ehsani, K., Lekkala, K., Ellis, K., Rana, K., Srinivasan, K., Fang, K., Singh, K.P., Zeng, K.H., Hatch, K., Hsu, K., Itti, L., Chen, L.Y., Pinto, L., Fei-Fei, L., Tan, L., Fan, L.J., Ott, L., Lee, L., Weihs, L., Chen, M., Lepert, M., Memmel, M., Tomizuka, M., Itkina, M., Castro, M.G., Spero, M., Du, M., Ahn, M., Yip, M.C., Zhang, M., Ding, M., Heo, M., Srirama, M.K., Sharma, M., Kim, M.J., Kanazawa, N., Hansen, N., Heess, N., Joshi, N.J., Suenderhauf, N., Liu, N., Di Palo, N., Shafiullah, N.M.M., Mees, O., Kroemer, O., Bastani, O., Sanketi, P.R., Miller, P.T., Yin, P., Wohlhart, P., Xu, P., Fagan, P.D., Mitrano, P., Sermanet, P., Abbeel, P., Sundaresan, P., Chen, Q., Vuong, Q., Rafailov, R., Tian, R., Doshi, R., Martín-Martín, R., Baijal, R., Scalise, R., Hendrix, R., Lin, R., Qian, R., Zhang, R., Mendonca, R., Shah, R., Hoque, R., Julian, R., Bustamante, S., Kirmani, S., Levine, S., Lin, S., Moore, S., Bahl, S., Dass, S., Sonawani, S., Song, S., Xu, S., Haldar, S., Karamcheti, S., Adebola, S., Guist, S., Nasiriany, S., Schaal, S., Welker, S., Tian, S., Ramamoorthy, S., Dasari, S., Belkhale, S., Park, S., Nair, S., Mirchandani, S., Osa, T., Gupta, T., Harada, T., Matsushima, T., Xiao, T., Kollar, T., Yu, T., Ding, T., Davchev, T., Zhao, T.Z., Armstrong, T., Darrell, T., Chung, T., Jain, V., Vanhoucke, V.,

Zhan, W., Zhou, W., Burgard, W., Chen, X., Wang, X., Zhu, X., Geng, X., Liu, X., Liangwei, X., Li, X., Lu, Y., Ma, Y.J., Kim, Y., Chebotar, Y., Zhou, Y., Zhu, Y., Wu, Y., Xu, Y., Wang, Y., Bisk, Y., Cho, Y., Lee, Y., Cui, Y., Cao, Y., Wu, Y.H., Tang, Y., Zhu, Y., Zhang, Y., Jiang, Y., Li, Y., Li, Y., Iwasawa, Y., Matsuo, Y., Ma, Z., Xu, Z., Cui, Z.J., Zhang, Z., Lin, Z.: Open x-embodiment: Robotic learning datasets and rt-x models : Open x-embodiment collaboration0. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 6892–6903 (2024). https://doi.org/10.1109/ICRA57147.2024.10611477

43. Qian, J., Li, Y., Bucher, B., Jayaraman, D.: Task-oriented hierarchical object decomposition for visuomotor control. In: Proceedings of The 8th Conference on Robot Learning. vol. 270, pp. 1891–1909 (2025), https://proceedings.mlr.press/v270/qian25b.html
44. Qian, J., Panagopoulos, A., Jayaraman, D.: Discovering deformable keypoint pyramids. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI. pp. 545–561 (2022). https://doi.org/10.1007/978-3-031-19809-0_31
45. Qian, J., Panagopoulos, A., Jayaraman, D.: Recasting generic pretrained vision transformers as object-centric scene encoders for manipulation policies. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 17544–17552 (2024). https://doi.org/10.1109/ICRA57147.2024.10610131
46. Rana, K., Lee, R., Pershouse, D., Suenderhauf, N.: Imle policy: Fast and sample efficient visuomotor policy learning via implicit maximum likelihood estimation. In: Proceedings of Robotics: Science and Systems (RSS) (2025). https://doi.org/10.15607/RSS.2025.XXI.158
47. Reuss, M., Pari, J., Agrawal, P., Lioutikov, R.: Efficient diffusion transformer policies with mixture of expert denoisers for multitask learning. In: International Conference on Learning Representations (2025)
48. Reuss, M., Yağmurlu, Ö.E., Wenzel, F., Lioutikov, R.: Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. In: Proceedings of Robotics: Science and Systems (RSS) (2024). https://doi.org/10.15607/RSS.2024.XX.121
49. Reuss, M., Zhou, H., Rühle, M., Yağmurlu, Ö.E., Otto, F., Lioutikov, R.: FLOWER: Democratizing generalist robot policies with efficient vision-language-flow models. In: Proceedings of The 9th Conference on Robot Learning. vol. 305, pp. 3736–3761 (2025), https://proceedings.mlr.press/v305/reuss25a.html
50. Ross, S., Gordon, G., Bagnell, J.: A reduction of imitation learning and structured prediction to no-regret online learning. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. vol. 15, pp. 627–635 (11–13 Apr 2011), https://proceedings.mlr.press/v15/ross11a.html
51. Schaal, S.: Learning from demonstration. In: Proceedings of the 10th International Conference on Neural Information Processing Systems. pp. 1040–1046 (1996)
52. Schaal, S., Peters, J., Nakanishi, J., Ijspeert, A.: Learning movement primitives. International Journal of Robotic Research - IJRR **15**, 561–572 (01 2003). https://doi.org/10.1007/11008941_60
53. Seitzer, M., Horn, M., Zadaianchuk, A., Zietlow, D., Xiao, T., Simon-Gabriel, C.J., He, T., Zhang, Z., Schölkopf, B., Brox, T., Locatello, F.: Bridging the gap to real-world object-centric learning. In: International Conference on Learning Representations (2022)
54. Shi, J., Qian, J., Ma, Y.J., Jayaraman, D.: Composing pre-trained object-centric representations for robotics from "what" and "where" foundation models. In: 2024

IEEE International Conference on Robotics and Automation (ICRA). pp. 15424–15432 (2024). https://doi.org/10.1109/ICRA57147.2024.10610695
55. Shridhar, M., Manuelli, L., Fox, D.: Cliport: What and where pathways for robotic manipulation. In: Proceedings of the 5th Conference on Robot Learning. vol. 164, pp. 894–906 (2022), https://proceedings.mlr.press/v164/shridhar22a.html
56. Shridhar, M., Manuelli, L., Fox, D.: Perceiver-actor: A multi-task transformer for robotic manipulation. In: Proceedings of The 6th Conference on Robot Learning. vol. 205, pp. 785–799 (2023), https://proceedings.mlr.press/v205/shridhar23a.html
57. Sieb, M., Xian, Z., Huang, A., Kroemer, O., Fragkiadaki, K.: Graph-structured visual imitation. In: Proceedings of the Conference on Robot Learning. vol. 100, pp. 979–989 (2020), https://proceedings.mlr.press/v100/sieb20a.html
58. Singh, G., Deng, F., Ahn, S.: Illiterate DALL-e learns to compose. In: International Conference on Learning Representations (2022)
59. Sun, J., Li, Y., Xu, L., Wang, N., Wei, J., Zhang, Y., Lu, C.: Conceptfactory: Facilitate 3d object knowledge annotation with object conceptualization. In: Proceedings of the 38th International Conference on Neural Information Processing Systems (2024)
60. Tatarchenko, M., Richter, S.R., Ranftl, R., Li, Z., Koltun, V., Brox, T.: What do single-view 3d reconstruction networks learn? In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3400–3409 (2019). https://doi.org/10.1109/CVPR.2019.00352
61. Todorov, E., Erez, T., Tassa, Y.: Mujoco: A physics engine for model-based control. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 5026–5033 (2012). https://doi.org/10.1109/IROS.2012.6386109
62. Torabi, F., Warnell, G., Stone, P.: Recent advances in imitation learning from observation. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. pp. 6325–6331 (2019). https://doi.org/10.24963/ijcai.2019/882
63. Tremblay, J., To, T., Sundaralingam, B., Xiang, Y., Fox, D., Birchfield, S.: Deep object pose estimation for semantic robotic grasping of household objects. In: Proceedings of The 2nd Conference on Robot Learning. vol. 87, pp. 306–316 (2018), https://proceedings.mlr.press/v87/tremblay18a.html
64. Tyree, S., Tremblay, J., To, T., Cheng, J., Mosier, T., Smith, J., Birchfield, S.: 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 13081–13088 (2022). https://doi.org/10.1109/IROS47612.2022.9981838
65. Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., Savarese, S.: Densefusion: 6d object pose estimation by iterative dense fusion. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3338–3347 (2019). https://doi.org/10.1109/CVPR.2019.00346
66. Wang, D., Devin, C., Cai, Q.Z., Yu, F., Darrell, T.: Deep object-centric policies for autonomous driving. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 8853–8859 (2019). https://doi.org/10.1109/ICRA.2019.8794224
67. Wang, Y., Yin, G., Huang, B., Kelestemur, T., Wang, J., Li, Y.: Gendp: 3d semantic fields for category-level generalizable diffusion policy. In: Proceedings of The 8th Conference on Robot Learning. vol. 270, pp. 4866–4878 (2025), https://proceedings.mlr.press/v270/wang25m.html

68. Wu, H., Jing, Y., Cheang, C., Chen, G., Xu, J., Li, X., Liu, M., Li, H., Kong, T.: Unleashing large-scale video generative pre-training for visual robot manipulation. In: International Conference on Learning Representations (2024)
69. Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., Levine, S.: Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In: Proceedings of the Conference on Robot Learning. vol. 100, pp. 1094–1100 (2020), https://proceedings.mlr.press/v100/yu20a.html
70. Yuan, W., Khot, T., Held, D., Mertz, C., Hebert, M.: Pcn: Point completion network. In: 2018 International Conference on 3D Vision (3DV). pp. 728–737 (2018). https://doi.org/10.1109/3DV.2018.00088
71. Yuan, W., Paxton, C., Desingh, K., Fox, D.: Sornet: Spatial object-centric representations for sequential manipulation. In: Proceedings of the 5th Conference on Robot Learning. vol. 164, pp. 148–157 (2022), https://proceedings.mlr.press/v164/yuan22a.html
72. Yue, Y., Wang, Y., Kang, B., Han, Y., Wang, S., Song, S., Feng, J., Huang, G.: Deer-vla: dynamic inference of multimodal large language models for efficient robot execution. In: Proceedings of the 38th International Conference on Neural Information Processing Systems (2024)
73. Zare, M., Kebria, P.M., Khosravi, A., Nahavandi, S.: A survey of imitation learning: Algorithms, recent developments, and challenges. IEEE Transactions on Cybernetics **54**(12), 7173–7186 (2024). https://doi.org/10.1109/TCYB.2024.3395626
74. Ze, Y., Zhang, G., Zhang, K., Hu, C., Wang, M., Xu, H.: 3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations. In: Proceedings of Robotics: Science and Systems (RSS) (2024)
75. Zhang, Q., Liu, Z., Fan, H., Liu, G., Zeng, B., Liu, S.: Flowpolicy: enabling fast and robust 3d flow-based policy via consistency flow matching for robot manipulation. In: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'25/IAAI'25/EAAI'25 (2025). https://doi.org/10.1609/aaai.v39i14.33617
76. Zhao, H., Jiang, L., Jia, J., Torr, P., Koltun, V.: Point transformer. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16239–16248 (2021). https://doi.org/10.1109/ICCV48922.2021.01595
77. Zhao, T.Z., Kumar, V., Levine, S., Finn, C.: Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In: Proceedings of Robotics: Science and Systems (July 2023)
78. Zhou, Q.Y., Park, J., Koltun, V.: Open3D: A modern library for 3D data processing. arXiv:1801.09847 (2018)
79. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX (2022). https://doi.org/10.1007/978-3-031-20077-9_21
80. Zhu, H., Wang, Y., Huang, D., Ye, W., Ouyang, W., He, T.: Point cloud matters: rethinking the impact of different observation spaces on robot learning. In: Proceedings of the 38th International Conference on Neural Information Processing Systems (2025)
81. Zhu, Y., Jiang, Z., Stone, P., Zhu, Y.: Learning generalizable manipulation policies with object-centric 3d representations. In: 7th Annual Conference on Robot Learning. vol. 229, pp. 3418–3433 (2023), https://proceedings.mlr.press/v229/zhu23b.html

82. Zhu, Y., Joshi, A., Stone, P., Zhu, Y.: Viola: Imitation learning for vision-based manipulation with object proposal priors. In: Proceedings of The 6th Conference on Robot Learning. vol. 205, pp. 1199–1210 (2023), https://proceedings.mlr.press/v205/zhu23a.html