Spotlighting Task-Relevant Features: Object-Centric Representations for Better Generalization in Robotic Manipulation

Alexandre Chapin[1], Bruno Machado[1], Emmanuel Dellandrea[1], Liming Chen[1]

***Abstract*— Training robotic policies that reliably generalize to novel environments remains a persistent challenge. While state-of-the-art models leverage powerful global or dense visual features, these representations struggle to separate critical task-specific signals from background noise, causing failures under visual shifts. In this work, we investigate *Slot-based Object-Centric Representations (SOCRs)* as a structural solution that decomposes scenes into discrete, actionable entities without supervision. Through a large-scale diagnostic study across simulated (METAWORLD, LIBERO) and real-world tasks, we systematically evaluate SOCRs against dominant baselines. We find that structured abstraction drives inherent robustness: SOCRs drastically outperform standard models under severe lighting, texture, and clutter shifts without task-specific fine-tuning, and they scale effectively with in-domain video pre-training. However, we also identify a critical vulnerability: a structural capacity trade-off that leads to *slot merging* under high clutter. By mapping out both the strengths and the bottleneck limits of SOCRs, this work provides a clear roadmap for integrating structured visual abstractions into next-generation, generalizable robotic systems. Project website: http://spotlight-iros.github.io/**

I. INTRODUCTION

A **visuomotor policy** tells a robot what to do based on what it sees given a goal, mapping raw visual inputs to motor actions. State-of-the-art approaches for visuomotor policy learning are data-driven through deep learning, *e.g.*, imitation learning (IL) [3], and leverage a few expert demonstrations without the need for explicit programming. As such, the visual representation of the robot visual input is a key to the robot's **generalization** capability when the learned policy must adapt to objects, environments, and tasks that differ from those encountered during training [3].

Recent work has intensified efforts to improve visual representation learning and their pre-training methodology as the foundation for robust robot policy learning. Advances include large-scale pre-training on egocentric video with time-contrastive and language-aligned objectives [4], [5], [6], distilling vision foundation models into compact encoders [7], and adopting self-supervised schemes such as masked autoencoding [8]. However, despite their diversity, these methods mostly produce one of two features types: **global** or **dense features**. **Global features** is a single vector summarizing the entire image, typically obtained by pooling operations (*e.g.*, max or average pooling in CNNs) or by extracting a special CLS token in ViTs while **dense features** is an embedding from one of the encoder's layers (usually the last). While effective for in-distribution tasks, these representations lack an explicit mechanism to separate task-relevant objects from irrelevant background noise, as illustrated in Figure 1. Consequently, the resulting policies often overfit to spurious correlations, such as table texture or lighting conditions, leading to catastrophic failure under distribution shifts [9], [10].

To address these limitations, we revisit a fundamental question in visuomotor control: *What structural properties must a visual representation satisfy in order to yield policies that generalize under distribution shift?*

We posit a **structural bottleneck hypothesis**: representations that explicitly decompose scenes into discrete, object-level entities impose an inductive bias that promotes invariance to low-level appearance changes and reduces reliance on spurious correlations. In contrast, global or dense feature encodings entangle task-relevant and irrelevant signals, forcing downstream policies to implicitly learn separation from mixed representations. Slot-based Object-Centric Representations (SOCRs) [11], [12] instantiate such a bottleneck through competitive attention over a fixed set of latent slots much in line with theories of human perception [13], [14]. While prior work has explored slot-based models in unsupervised perception or simplified synthetic control settings [15], [16], [17], [18], [19], these studies have primarily emphasized **relational complexity**, focusing on multi-object reasoning and physical interactions within visually simplified environments. Consequently, the axis of **perceptual complexity** remains largely unexplored. It remains unclear whether (i) structured bottlenecks scale to the visual noise and distractors inherent to real-world robotic manipulation, (ii) they provide measurable robustness gains over modern foundation models (*e.g.*, DINOv2 [20], Theia [7]), and (iii) what structural limitations govern their behavior under increasing scene complexity.

In this work, we move beyond performance benchmarking and provide the first systematic diagnostic study of SOCRs for robot policy learning. We analyze how representation structure impacts generalization, identify concrete failure modes arising from capacity constraints, and provide guidance on how to overcome problems grounded in this structural understanding.

Our contributions are fourfold:

- **Validating structural robustness.** We empirically validate that SOCRs inherently improve robustness against different distribution shifts across simulated and real-world manipulation tasks.

[1]Ecole Centrale de Lyon, CNRS, Universite Claude Bernard Lyon 1, INSA Lyon, Université Lumière Lyon 2, LIRIS, UMR5205, 69130 Ecully, France

Correspondance to : alexandre.chapin@ec-lyon.fr

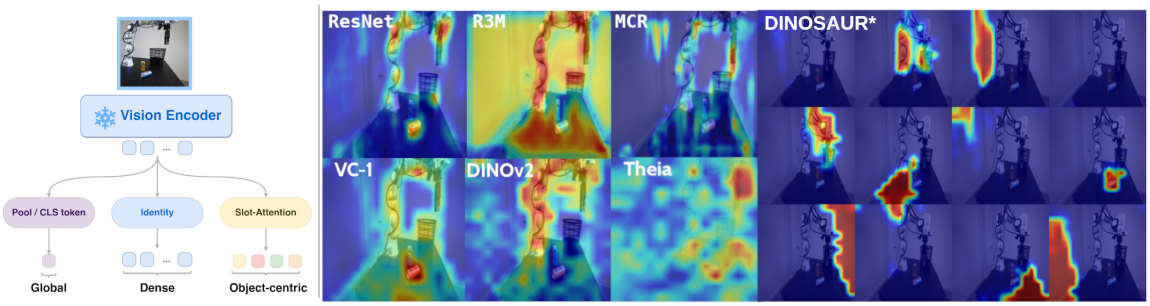

Fig. 1. **Overview of visual representations.** (*Left*) We use a set of pre-trained visual models with different latent-space structures: global, dense, and object-centric. Dense representations are extracted from one of the encoder's layers (CNN or ViT) before linear projection, while global representations are obtained after pooling operations. Slot-based object-centric representations emerge from an additional Slot-Attention layer that binds every dense feature to a finite set of slots. (*Right*) We visualize how each representation attends to different parts of the image (using GradCAM [1] for CNN-based methods and Attention-Rollout [2] for ViT-based). Most methods focus narrowly and may be distracted by irrelevant regions. In contrast, object-centric representations (DINOSAUR*) attend to multiple parts, naturally separating task-relevant from irrelevant information.

- **Benefits of large-scale pre-training.** We demonstrate that, contrary to existing assumptions [21], object-centric methods can effectively leverage large-scale pre-training to significantly boost downstream performance.
- **Identifying key failure modes.** We provide the first systematic breakdown of why object-centric representations fail in downstream control, pinpointing slot merging and capacity limits as the primary bottlenecks.
- **Balancing capacity and robustness.** Through targeted slot-number ablations, we reveal that structural capacity directly dictates out-of-distribution (OOD) robustness.

II. Related works

a) Pretrained vision-based models for robot learning: Large-scale vision models (MoCo [22], DINO [23], DINOv2 [20], CLIP [24]) have proven surprisingly effective for visuomotor policy learning [25]. To improve alignment with manipulation tasks, subsequent works introduced domain-specific pretraining using egocentric or robotic videos (R3M [4], [26], VC-1 [5], [27], Theia [7], MCR [28]). However, recent studies demonstrate that dataset quality, diversity, and emergent object-segmentation properties often outweigh strict domain alignment for generalization [9], [29]. These findings motivate our shift toward structured, object-centric representations that naturally capture task-relevant scene composition.

b) Slot-based object-centric representations: Slot-based object-centric representations (SOCRs) decompose visual scenes into structured latent entities (slots) in an unsupervised manner, gaining traction in autonomous driving, robotics, and explainability [16], [30], [31], [32]. Originating from generative models [18], [33], [34], methods like Slot Attention [11] have evolved to incorporate advanced decoders [35], [36], [37], pretrained backbones [12], and temporal dynamics for video [19], [38], [39], [40]. While SOCRs show potential for aiding generalization [9], [28], their application in control or imitation learning remains largely limited to simple, synthetic environments and are evaluated on in-domain scenarios only [15], [16], [17]. Crucially, these existing approaches fail to assess how well SOCRs generalize to new distribution scenarios, such as visual distractors, novel object configurations, or altered backgrounds. Without rigorous out-of-distribution (OOD) testing, it remains unclear whether the theoretical compositionality of slots translates into robust, real-world control policies under environmental shifts. To address this critical gap, our work systematically evaluates SOCRs in realistic robotic manipulation tasks under distributional shifts, comparing them against state-of-the-art dense visual representations.

c) Segmentation-driven object decomposition for robotics: An alternative approach pairs supervised foundation models like SAM [41] with frozen vision backbones to generate object masks (e.g., POCR [42], GROOT [43], HODOR [44]). While effective, these methods face key limitations: they rely on heavily annotated pretraining datasets, are memory and compute intensive (hindering real-time application), and often require explicit spatial prompts like bounding boxes. Conversely, SOCRs learn object decomposition end-to-end without supervision, offering a more flexible and computationally efficient alternative for real-time robotic inference.

III. Method

We first introduce the process of extracting object-centric representations from dense features. Then, we detail the integration of the object-centric visual features into a policy training framework.

a) Object-centric representation: Given an input image I, the goal is to produce a set of K object representations, or slots, $S = s_1, \ldots, s_K$. To achieve this, the image is first encoded by a visual backbone into dense feature tokens $F = f_1, \ldots, f_N$, where $N \gg K$. Slot Attention [11] then extracts a compact set of object representations S by iteratively attending to these features. It is a differentiable module that performs iterative cross-attention with competition, encouraging different slots to specialize in distinct parts of the input image. Formally, attention weights are computed as:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}}\right), \quad S^{(i+1)} = \mathbf{A}\mathbf{V}, \tag{1}$$

where the queries $\mathbf{Q}$ are projections of the current slot representations $S^{(i)}$, and the keys $\mathbf{K}$ and values $\mathbf{V}$ are projections of the feature tokens F. D is the projected feature dimension. This iterative refinement process yields the final object-centric slots S.

The SOCR model we used, which we refer to as DINOSAUR*, builds upon DINOSAUR [12] while introducing **two key modifications** to enhance representation quality and temporal consistency. First, we modernize the visual backbone by replacing the original DINO [23] encoder with the more robust DINOv2 [20]. Like the original model, we utilize a decoder to reconstruct these high-level backbone features rather than raw pixels. Second, we extend the static Slot-Attention mechanism to the temporal domain. Inspired by [38], [40], we incorporate a Transformer layer between timesteps to facilitate recursive information transfer. In this setup, slots at each timestep first undergo slot-wise self-attention within the Transformer layer to aggregate context; these refined slots then serve as the initialization for the subsequent Slot-Attention module.

b) Policy training: Given a dataset of expert demonstrations $\mathcal{D} = \tau_1, \ldots, \tau_n$, where each trajectory $\tau_i = [(o_0, a_0), \ldots, (o_T, a_T)]$ pairs observations with actions, we aim to learn a policy π that learns to map the current observation o_t (e.g. visual inputs) to the next action a_t. To ensure a fair comparison across different feature types, we utilize a framework inspired by [45]. It consists of an encoder, an observation trunk, and a policy head. The observation trunk is a transformer encoder that encodes a sequence of T past observations, each containing visual features, proprioceptive states, and language embeddings, interleaved with learnable action tokens. The policy head, a MLP, generates the next action from the action tokens. The architecture treats visual inputs as sets of tokens, this property ensures that the transformer can attend over both structured (slot-based) and unstructured (global or dense) features, thereby preserving fairness across representation types . Importantly, each policy is trained in a **multi-task setting**, enabling learning from demonstrations across multiple tasks rather than a single one. All pretrained vision encoders are **kept frozen** during policy training to isolate the effect of representation choice.

c) Robotic pre-training: While object-centric models excel at structured scene decomposition, they are primarily trained on *in-the-wild* datasets. This can result in distribution shift when these models are applied to robotic manipulation environments. To bridge this gap and enhance data diversity, aligning with the observations in [29], we introduce a dedicated domain-specific pretraining stage using large-scale robotic video data. To rigorously evaluate the impact of this domain alignment, we develop two versions of our model: **DINOSAUR***, trained on the standard COCO dataset for fair comparison against existing baselines, and **DINOSAUR-Rob***, which is pre-trained specifically on robotic data. Our robotic pre-training corpus is a weighted combination of real-world robotic datasets [46], [47], [48] spanning a vast array of manipulation skills, environments, and robotic embodiments.

IV. Benchmarks and Experimental setup

a) Environments and tasks: To comprehensively evaluate visual representations for robot manipulation, we selected **three environments**, two in simulation and one in the real world, chosen for their diversity in task complexity, embodiment, and visual structure.

In simulation, we use MetaWorld [49], a widely adopted benchmark consisting of tabletop manipulation tasks with a Sawyer robotic arm. MetaWorld provides a controlled, standardized setting, making it ideal for evaluating representations in simple, single-object scenarios and testing basic generalization (distractors, texture, and lighting). We use the same task set as in [28].

To assess multi-object reasoning, we also include LIBERO [50], a recent benchmark featuring scenes across kitchens, offices, and living rooms. We train and evaluate on the LIBERO-90 tasks, which involve multiple objects with diverse appearances and affordances, emphasizing combinatorial generalization and reasoning about object interactions. However, LIBERO's benchmark does not introduce any distributional shifts for evaluation. So we only report results on the training distribution.

For real-world evaluation, we deploy a Franka robotic arm on a set of four tabletop manipulation tasks, Figure 2 provides an overview of the setup used. For each task, we collected 50 expert demonstrations via teleoperation. Task success is computed using success rate. A trial is a success if the final physical configuration meets the task-specific criteria: the bowls are nested inside the pan, the screwdriver is contained within the closed drawer, all cans are positioned inside the bin, or the plates are inside the slots of the dish rack.

b) Pre-trained visual models: We selected seven representative pre-trained visual encoders, detailed in Table II. This selection spans a diverse architectural landscape, including ResNet variants [51], Vision Transformers (ViTs) [52], and encompasses a variety of training objectives such as supervised learning, self-supervised learning, contrastive learning, and distillation. While these models utilize different pre-training datasets, we follow the findings of Parisi et al. [25], which suggest that the choice of pre-training data is often secondary to the representation architecture for downstream control tasks. Our benchmark aims to capture all major feature representation types to facilitate a comprehensive evaluation: **global** features (a single token representing the entire image), **dense** features (a grid of tokens capturing spatial detail), and **slot-based** features (object-centric representations that decompose the scene into discrete entities). For transparency regarding computational complexity, Table II reports the total number of tokens provided as input to the policy for each model. Although models like ResNet-50 and DINOv2 support both Global and Dense features, we report all results using their Dense representations, as these consistently yielded superior performance in both in-domain and out-of-domain evaluations. Additionally, our comparison includes a segmentation-driven object-centric baseline which

Fig. 2. **Overview of real-world setup.** We evaluate the different visual models on a Franka robotic arm on four tabletop manipulation tasks (From left to right): Stacking bowls into a pan, Opening a drawer placing a screwdriver inside and closing the drawer, Putting cans into a bin and Placing plates into a dish rack.

integrates the Segment Anything Model (SAM) [41] with DINOv2 [20] inspired by [42].

c) Impact of robotic pre-training: To understand the impact of domain alignment, Table I ablates the choice of different robotic datasets used for pre-training. A mixture of diverse robotic datasets not only improves performance over using any individual robotic dataset, but also outperforms the standard vision baseline (COCO) across all evaluation domains.

TABLE I

PRE-TRAINING DATA EVALUATION. FINAL GLOBAL PERFORMANCE FOR DINOSAUR*.

Pre-training data	MetaWorld	LIBERO	Real-Robot
BridgeV2 (B) [47]	0.75	0.73	0.45
Fractal (F) [48]	0.72	0.58	0.38
DROID (D) [46]	0.74	0.72	0.26
DINOSAUR-Rob* (D+B+F)	**0.76**	**0.77**	**0.56**

V. RESULTS

We evaluate the role of different visual features types in learning and generalizing robotic manipulation policies. Our work aim to answer the following questions:

- **Q1:** Can SOCRs improve robot policy learning efficiency over other visual representations?
- **Q2:** Can SOCRs enhance policy generalization under visual distribution shifts ?
- **Q3:** When do SOCRs fail, and what factors impacts their performance ?

In simulation, we report mean success rates over three random seeds with 50 rollouts per task; in the real world, we conduct 12 rollouts per task and per generalization level ($\sim$1000 rollouts in total by combining all experiments).

A. Q1: Do SOCRs improve manipulation policy learning ?

Figure 3 summarizes the average success rate across Meta-World, LIBERO, and our real-world benchmark. The green dots represent in-domain performance, while the orange dots indicate average success rates across various generalization scenarios (distractors, novel textures, lighting changes). The red numbers denote the relative drop in performance from in-domain to generalization settings. Policy models based on object-centric features, especially DINOSAUR-Rob*, consistently achieve the highest overall performance, on par with or surpassing the best dense and global baselines across all environments. This confirms that SOCRs not only achieve effective policy learning but also scale well to complex, multi-object scenarios.

We observe that the segmentation-driven representation (SAM+DINOv2) prevents learning effective policies. Indeed, contrary to prior work [42], [43], we use a version only encoding the "what" (appearance) of objects, without any spatial information (bounding boxes or masks coordinates). This choice was made to ensure a fair comparison with other models, as they do not have access to such spatial cues. However, this design decision likely hampers the model's ability to capture object relationships and spatial arrangements, which are crucial for effective manipulation. This also highlights the limitations of segmentation-driven approaches compared to end-to-end learned object-centric representations that can capture both appearance and spatial structure jointly during training.

In **MetaWorld**, all models except the VC-1-based perform above 60%. The low VC-1 performance may result from its MAE-based pretraining, which is sensitive to domain mismatch when not fine-tuned. Object-centric-based models perform comparably to top baselines here, despite the simplicity of the environment.

In **LIBERO**, which features complex scenes with multiple objects, SOCRs-based policies perform on par with or better than dense-based (Theia, DINOv2). The global-based models (R3M, VC-1) lag behind, likely due to their inability to capture fine-grained object interactions.

In the **real-world**, the ResNet-based policy is surprisingly strong, likely due to the diversity of ImageNet pretraining, consistent with [29]. The DINOSAUR-Rob* policy performs on par with ResNet and outperforms all other baselines by a significant margin, achieving a 56% success rate. Notably, even without robotic pretraining, the DINOSAUR* policy remains highly competitive, achieving a 48% success rate on real-world tasks and improve 20% over the DINOv2 policy. This suggests that the intermediate object-centric structuration itself confers robustness and effectiveness.

Overall, our results for Q1 confirm that **object-centric models are not only effective but scalable across domains**, offering performance benefits in both structured simulation tasks and noisy real-world environments.

B. Q2: Do SOCRs enhance generalization under visual distribution shifts?

We now evaluate generalization to out-of-distribution conditions, including novel distractors, unseen textures, and lighting changes. Table III, and IV provide detailed results

TABLE II

MODELS OVERVIEW. COMPARISON OF THE DATA AND SIZES OF THE DIFFERENT PRE-TRAINED VISUAL MODELS . VIT: VISION TRANSFORMER, SOCR: SLOT-BASED OBJECT-CENTRIC LAYERS, G: GLOBAL, D: DENSE, SM: SEGMENTATION.

Model	Backbone	Pre-training Dataset	# of params.	Features	# of tokens
DINOv2 [20]	ViT	LVD-142M [20]	86M	D/G	196/1
VC-1 [5]	ViT	Ego4D [26] and ImageNet [27]	86M	G	1
Theia [7]	ViT	ImageNet [27]	140M	D	196
R3M [4]	ResNet-50	Ego4D [26]	25.6M	G	1
ResNet-50 [51]	ResNet-50	ImageNet [27]	25.6M	D/G	49/1
SAM [41] + DINOv2	ViT + SM	LVD-142M [20] + SA-V [41]	∼400M	SM	10
DINOSAUR*	ViT + SOCR	COCO [53]	88M	Slot	10
DINOSAUR-Rob*	ViT + SOCR	Robot Mixt.	88M	Slot	10

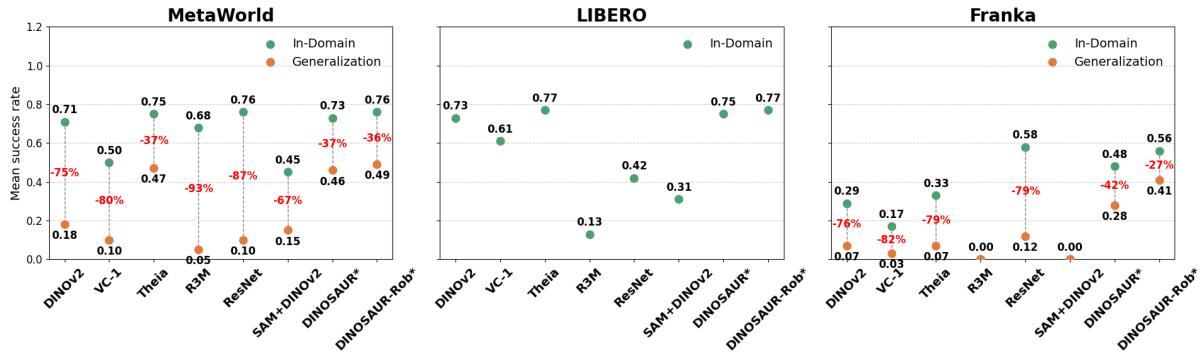

Fig. 3. **Overall success rate on in-domain and generalization scenarios.** Mean success rate over all tasks for each visual model on MetaWorld (left), LIBERO (middle) and Real robot using Franka (right). Green dot: in-domain performance, Orange dot: average performance over all generalization scenarios (distractors, novel textures, lighting changes). Red number: relative drop in performance from in-domain to generalization settings. As LIBERO's benchmark does not introduce any distributional shifts for evaluation, we only report in-domain performances.

for each environment and shift type. As SAM+DINOv2 fails to learn effective policies, we exclude it from the generalization analysis.

Overall, SOCRs-based policies exhibit substantially better robustness to distribution shifts compared to dense and global representations. Notably, DINOSAUR-Rob* policy consistently achieves the highest success rates overall across environments, with the smallest relative performance drop from in-domain to generalization settings (Figure 3). Even without robotic pretraining, DINOSAUR* remains highly competitive in simulation and clearly outperforms all baselines in the real world. Notably, we can have a direct comparison between DINOSAUR* and DINOv2 as they share the same backbone with only the additional adapted Slot-Attention layer on top of the frozen model. This comparison highlights the benefits of object-centric intermediate representations for generalization, as DINOSAUR* significantly outperforms DINOv2 across all scenarios.

In **MetaWorld** (Table III), DINOv2 and Theia-based policies outperform ResNet-based policies on average, confirming previous findings (e.g.,[9]). Notably, Theia performs best in the distractor scenario, likely due to CLIP-based text-image alignment helping the policy to ignore irrelevant patches. Remarkably, the SOCRs-based models excel under texture and lighting shifts, where they outperform all baselines by a large margin. This suggests that object-centric representations effectively capture invariant object properties, enhancing robustness to appearance changes. Indeed,

TABLE III

GENERALIZATION METAWORLD. COMPARISON OF DIFFERENT LEVELS OF GENERALIZATION IN METAWORLD.

Models	Distractors	Textures	Lighting	Over.
ResNet-50	0.04 ± 0.02	0.0 ± 0.0	0.22 ± 0.02	0.10
R3M	0.0	0.0	0.0	0.0
DINOv2	0.11 ± 0.02	0.03 ± 0.01	0.39 ± 0.04	0.18
VC1	0.06 ± 0.02	0.0 ± 0.0	0.23 ± 0.03	0.10
Theia	**0.65** ± 0.10	0.28 ± 0.06	0.48 ± 0.10	0.47
DINOSAUR*	0.21 ± 0.04	**0.48** ± 0.06	**0.71** ± 0.06	0.46
DINOSAUR-Rob*	0.46 ± 0.14	0.36 ± 0.05	0.65 ± 0.14	**0.49**

Figure 1 shows that SOCRs can distinguish objects from background clutter, likely enabling policies to focus on task-relevant elements.

In **real-world evaluations** (Table IV), both SOCRs-based models excels in the two generalization levels and overperform by a large margin every other features types. The ResNet-50-based policy, which performed competitively in-domain, struggles significantly under distribution shifts, likely due to its limited capacity to disentangle objects from background noise. Dense-based models also see substantial performance drops, underscoring their sensitivity to visual perturbations. In contrast, SOCRs-based models maintain robust performance, with DINOSAUR-Rob* achieving a 41% success rate on average across shifts, highlighting the practical benefits of object-centric representations for real-world robotic manipulation.

In summary, results for Q2 demonstrate that **object-**

TABLE IV

GENERALIZATION REAL-WORLD. COMPARISON OF DIFFERENT LEVELS OF GENERALIZATION IN THE REAL ROBOT.

Models	Distractors	Textures	Overall
ResNet-50	0.15	0.10	0.12
R3M	0.0	0.0	0.0
DINOv2	0.06	0.08	0.07
VC1	0.03	0.02	0.03
Theia	0.06	0.08	0.07
DINOSAUR*	0.27	0.29	0.28
DINOSAUR-Rob*	**0.37**	**0.44**	**0.41**

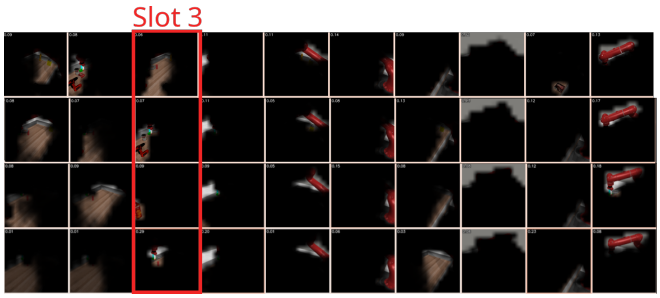

Fig. 4. **Qualitative analysis of distractor generalization failures (K=10).** We visualize the slot decomposition produced by the DINOSAUR* model across four difficulty levels in MetaWorld. From top to bottom: *Hard*, *Medium*, and *Easy* distractor scenarios, and *In-Domain* baseline (bottom row). Ideally, the gripper and object (Slot 3 in the baseline) should remain isolated. However, we observe *slot merging* in the different levels of clutter.

centric representations generalize better across diverse distribution shifts, particularly those that perturb low-level appearance. This robustness likely stems from SOCRs ability to filter task-irrelevant background and focus on object-level structure.

C. Q3: When do SOCRs fail, and what factors impacts their performance ?

Even though object-centric models are more robust to distractors than most methods, there is still a significant performance drop compared to other distribution shifts. To diagnose the root cause of this performance degradation, we analyze the slot attention maps across varying levels of visual distraction (Figure 4). While the slots capturing the background and the robot arm remain largely unaffected by the shift, we identify a key phenomena that degrades representation quality: **slot merging**. This occurs when features from distractor objects leak into task-relevant slots (e.g., those containing the target object or the gripper), effectively polluting critical state information with irrelevant visual noise.

These findings indicate that standard SOCRs, constrained by a fixed-capacity bottleneck, struggle to separate novel objects from task-relevant ones when objects outnumber slots. Consequently, a critical question arises regarding mitigation: Does increasing the number of slots resolve these overcrowding effects?

a) Impact of slot number: A distinct advantage of SOCR models is their flexibility to modify the number of slots (K) dynamically at inference time, as slots are initialized from a shared learnable Gaussian distribution [11]. We leverage this property to investigate whether modulating the information bottleneck can mitigate distractor interference, and avoid *slot merging*. Keeping the pre-trained DINOSAUR-Rob* model frozen, we extract representations using different slot capacities $K \in \{4, 7, 10, 14, 16, 20\}$ and train separate versions of the policy for each configuration. We evaluate these models across both In-Domain (ID) and Distractor (OOD) scenarios in MetaWorld. The results, illustrated in Figure 5, reveal an interesting trade-off between peak ID performance and OOD robustness depending on the slot capacity:

- **Under-segmentation ($K < 10$):** We observe a sharp performance collapse when reducing capacity, particularly at $K = 4$ where the OOD success rate drops to 34%. This confirms the "slot merging" failure mode identified previously: when K is lower than the number of distinct entities in the scene (robot, target, goal, plus distractors), the model is forced to compress distractors and task-relevant objects into shared slots, rendering the state representation ambiguous for the policy.
- **Increased Capacity ($K > 10$):** Increasing the slot budget to $K = 14, 16$, or 20 demonstrates a clear trade-off. While In-Domain performance decreases (dropping from 0.76 at $K = 10$ to ~ 0.61 at higher capacities), the robustness to OOD distractors notably improves, peaking at 0.60 for $K = 20$. The extra slots allow the visual encoder to isolate novel distractors into their own representations rather than merging them with task-relevant features. However, when there are less objects on the scene (ID, no distractors), the extra slots might decompose relevant objects into multiple parts, which can make the representation less useful for the policy, leading to a decline in ID success rate.

In summary, higher slot budgets ensure the visual encoder has enough bandwidth to allocate individual slots to novel OOD distractors without corrupting the representations of the robot or the target. By keeping these task-relevant features isolated, the fundamental state representation remains stable, which is why OOD performance actually climbs and stabilizes around 0.60 at higher capacities. However, passing a larger set of slots to the downstream policy introduces a new optimization challenge: the policy must now learn to identify and act upon the correct subset of objects out of a noisier set of vectors that represents objects and parts. The consistent degradation of ID performance (dropping from 0.76 to roughly 0.61) suggests that the policy architecture struggles to efficiently filter through this increased dimensionality. Ultimately, these findings indicate that while providing a larger slot budget is an effective strategy for avoiding distractor interference at the representation level, it is not a complete solution.

To address these limitations, we propose **two directions** for next-generation SOCR architectures. **Multi-granular representations**: models should utilize fine-grained slots

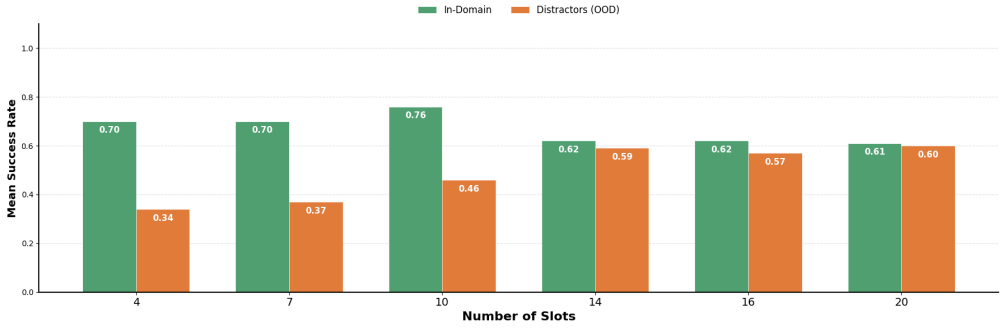

Fig. 5. **Slot count ablation.** We evaluate the impact of varying the number of slots (K) at inference using a DINOSAUR-Rob* model trained with $K = 10$. Reducing capacity ($K < 10$) exacerbates sensitivity to distractors, dropping OOD success rates significantly (to 0.34 at $K = 4$) due to forced slot merging. Conversely, increasing capacity ($K > 10$) improves OOD robustness (reaching up to 0.60 at $K = 20$) but causes a notable decline in In-Domain (ID) performance, highlighting a trade-off between representation stability against distractors and policy optimization complexity.

to capture visual complexity and coarse-grained slots to aggregate semantic entities, decoupling visual noise from policy inputs. **Adaptive capacity and Semantics**: dynamic slot capacity must be paired with selection mechanisms to distinguish active novel objects from passive background noise [54], [55]. Furthermore, future work must bridge the gap between unsupervised grouping and semantic understanding [56], ensuring that slots align with functional categories (e.g., "tool," "obstacle") rather than merely visual clusters.

VI. Conclusion

We formalized and empirically validated the structural bottleneck hypothesis for visuomotor control, showing that object-level abstraction, not backbone scale alone, drives robustness under distribution shift. We show that SOCR-based policies (specifically DINOSAUR-Rob*) consistently outperform traditional dense and global baselines. These benefits are most pronounced under severe distribution shifts in lighting and texture, confirming that object-centric inductive biases better reflect the structured nature of physical interaction than pixel-level features.

However, our analysis also highlights that robustness is not automatic. We observed that while SOCRs resist appearance shifts, they remain sensitive to high-clutter scenarios. Ablation studies on slot capacity (K) demonstrate a fundamental trade-off between representation fidelity and policy optimization. Specifically, a low capacity ($K < 10$) leads to under-segmentation, corrupting the state representation by merging distractor features into task-relevant slots. Conversely, a high capacity ($K > 10$) preserves the representation by isolating distractors, but it degrades in-domain performance by flooding the downstream policy with a noisy, overloaded state space that the architecture struggles to filter.

Our findings suggest that the path toward generalizable robotics lies in moving away from monolithic pixel features toward structured, object-based encodings. By resolving the stability and capacity trade-offs identified in this work, we can bridge the gap between low-level visual input and high-level symbolic reasoning, enabling robots to interact with dynamic, real-world environments with greater reliability.

Acknowledgment

This work was in part supported by the French Research Agency, l'Agence Nationale de Recherche (ANR), through the projects Chiron (ANR-20-IADJ-0001-01), Astérix (ANR-23-EDIA-0002) and the French national investment priority program PSPC FAIR WASTE project. It was granted access to the HPC resources of IDRIS under the allocation 2025-[AD011016842], 2025-[AD011016615] and 2026-[A0201017513] made by GENCI.

References

[1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, p. 336–359, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1007/s11263-019-01228-7

[2] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," 2020. [Online]. Available: https://arxiv.org/abs/2005.00928

[3] Z. Li, A. Chapin, E. Xiang, R. Yang, B. Machado, N. Lei, E. Dellandrea, D. Huang, and L. Chen, "Robotic manipulation via imitation learning: Taxonomy, evolution, benchmark, and challenges," 2025. [Online]. Available: https://arxiv.org/abs/2508.17449

[4] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," 2022. [Online]. Available: https://arxiv.org/abs/2203.12601

[5] A. Majumdar, K. Yadav, S. Arnaud, *et al.*, "Where are we in the search for an artificial visual cortex for embodied intelligence?" 2024. [Online]. Available: https://arxiv.org/abs/2303.18240

[6] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "Vip: Towards universal visual reward and representation via value-implicit pre-training," 2023. [Online]. Available: https://arxiv.org/abs/2210.00030

[7] J. Shang, K. Schmeckpeper, B. B. May, M. V. Minniti, T. Kelestemur, D. Watkins, and L. Herlant, "Theia: Distilling diverse vision foundation models for robot learning," 2024. [Online]. Available: https://arxiv.org/abs/2407.20179

[8] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," 2022. [Online]. Available: https://arxiv.org/abs/2210.03109

[9] K. Burns, Z. Witzel, J. I. Hamid, T. Yu, C. Finn, and K. Hausman, "What makes pre-trained visual representations successful for robust manipulation?" 2023. [Online]. Available: https://arxiv.org/abs/2312.12444

[10] Y. Hu, R. Wang, L. E. Li, and Y. Gao, "For pre-trained vision models in motor control, not all policy learning methods are created equal," 2023. [Online]. Available: https://arxiv.org/abs/2304.04591

[11] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," 2020. [Online]. Available: https://arxiv.org/abs/2006.15055

[12] M. Seitzer, M. Horn, A. Zadaianchuk, *et al.*, "Bridging the gap to real-world object-centric learning," 2023. [Online]. Available: https://arxiv.org/abs/2209.14860

[13] E. S. Spelke, "Principles of object perception," *Cognitive science*, vol. 14, no. 1, pp. 29–56, 1990.

[14] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," 2016. [Online]. Available: https://arxiv.org/abs/1604.00289

[15] J. Yoon, Y.-F. Wu, H. Bae, and S. Ahn, "An investigation into pre-training object-centric representations for reinforcement learning," 2023. [Online]. Available: https://arxiv.org/abs/2302.04419

[16] N. Heravi, A. Wahid, C. Lynch, P. Florence, T. Armstrong, J. Tompson, P. Sermanet, J. Bohg, and D. Dwibedi, "Visuomotor control in multi-object scenes using object-aware representations," 2023. [Online]. Available: https://arxiv.org/abs/2205.06333

[17] D. Haramati, T. Daniel, and A. Tamar, "Entity-centric reinforcement learning for object manipulation from pixels," 2024. [Online]. Available: https://arxiv.org/abs/2404.01220

[18] N. Watters, L. Matthey, M. Bosnjak, C. P. Burgess, and A. Lerchner, "Cobra: Data-efficient model-based rl through unsupervised object discovery and curiosity-driven exploration," 2019. [Online]. Available: https://arxiv.org/abs/1905.09275

[19] T. Kipf, G. F. Elsayed, A. Mahendran, A. Stone, S. Sabour, G. Heigold, R. Jonschkowski, A. Dosovitskiy, and K. Greff, "Conditional object-centric learning from video," 2022. [Online]. Available: https://arxiv.org/abs/2111.12594

[20] M. Oquab, T. Darcet, T. Moutakanni, *et al.*, "Dinov2: Learning robust visual features without supervision," 2024. [Online]. Available: https://arxiv.org/abs/2304.07193

[21] A. Didolkar, A. Zadaianchuk, A. Goyal, M. Mozer, Y. Bengio, G. Martius, and M. Seitzer, "Zero-shot object-centric representation learning," 2024. [Online]. Available: https://arxiv.org/abs/2408.09162

[22] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020. [Online]. Available: https://arxiv.org/abs/2003.04297

[23] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," 2021. [Online]. Available: https://arxiv.org/abs/2104.14294

[24] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[25] S. Parisi, A. Rajeswaran, S. Purushwalkam, and A. Gupta, "The unsurprising effectiveness of pre-trained vision models for control," 2022. [Online]. Available: https://arxiv.org/abs/2203.03580

[26] K. Grauman, A. Westbury, E. Byrne, *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," 2022. [Online]. Available: https://arxiv.org/abs/2110.07058

[27] O. Russakovsky, J. Deng, H. Su, *et al.*, "Imagenet large scale visual recognition challenge," 2015. [Online]. Available: https://arxiv.org/abs/1409.0575

[28] G. Jiang, Y. Sun, T. Huang, H. Li, Y. Liang, and H. Xu, "Robots pre-train robots: Manipulation-centric robotic representation from large-scale robot dataset," *arXiv preprint arXiv:2410.22325*, 2024.

[29] S. Dasari, M. K. Srirama, U. Jain, and A. Gupta, "An unbiased look at datasets for visuo-motor pre-training," 2023. [Online]. Available: https://arxiv.org/abs/2310.09289

[30] S. Hamdan and F. Güney, "Carformer: Self-driving with learned object-centric representations," 2024. [Online]. Available: https://arxiv.org/abs/2407.15843

[31] M. Mosbach, J. N. Ewertz, A. Villar-Corrales, and S. Behnke, "Sold: Slot object-centric latent dynamics models for relational manipulation learning from pixels," 2025. [Online]. Available: https://arxiv.org/abs/2410.08822

[32] B. Wang, L. Li, J. Zhang, Y. Nakashima, and H. Nagahara, "Explainable image recognition via enhanced slot-attention based classifier," 2024. [Online]. Available: https://arxiv.org/abs/2407.05616

[33] R. Kabra, D. Zoran, G. Erdogan, L. Matthey, A. Creswell, M. Botvinick, A. Lerchner, and C. P. Burgess, "Simone: View-invariant, temporally-abstracted object representations via unsupervised video decomposition," 2021. [Online]. Available: https://arxiv.org/abs/2106.03849

[34] C. P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, and A. Lerchner, "Monet: Unsupervised scene decomposition and representation," 2019. [Online]. Available: https://arxiv.org/abs/1901.11390

[35] J. Jiang, F. Deng, G. Singh, and S. Ahn, "Object-centric slot diffusion," 2023. [Online]. Available: https://arxiv.org/abs/2303.10834

[36] Z. Wu, J. Hu, W. Lu, I. Gilitschenski, and A. Garg, "Slotdiffusion: Object-centric generative modeling with diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2305.11281

[37] G. Singh, F. Deng, and S. Ahn, "Illiterate dall-e learns to compose," 2022. [Online]. Available: https://arxiv.org/abs/2110.11405

[38] G. F. Elsayed, A. Mahendran, S. van Steenkiste, K. Greff, M. C. Mozer, and T. Kipf, "Savi++: Towards end-to-end object-centric learning from real-world videos," 2022. [Online]. Available: https://arxiv.org/abs/2206.07764

[39] G. Singh, Y.-F. Wu, and S. Ahn, "Simple unsupervised object-centric learning for complex and naturalistic videos," 2022. [Online]. Available: https://arxiv.org/abs/2205.14065

[40] A. Zadaianchuk, M. Seitzer, and G. Martius, "Object-centric learning for real-world videos by predicting temporal feature similarities," 2023. [Online]. Available: https://arxiv.org/abs/2306.04829

[41] A. Kirillov, E. Mintun, N. Ravi, *et al.*, "Segment anything," 2023. [Online]. Available: https://arxiv.org/abs/2304.02643

[42] J. Shi, J. Qian, Y. J. Ma, and D. Jayaraman, "Composing pre-trained object-centric representations for robotics from "what" and "where" foundation models," 2024. [Online]. Available: https://arxiv.org/abs/2404.13474

[43] Y. Zhu, Z. Jiang, P. Stone, and Y. Zhu, "Learning generalizable manipulation policies with object-centric 3d representations," 2023. [Online]. Available: https://arxiv.org/abs/2310.14386

[44] J. Qian, Y. Li, B. Bucher, and D. Jayaraman, "Task-oriented hierarchical object decomposition for visuomotor control," 2024. [Online]. Available: https://arxiv.org/abs/2411.01284

[45] S. Haldar, Z. Peng, and L. Pinto, "Baku: An efficient transformer for multi-task policy learning," 2024. [Online]. Available: https://arxiv.org/abs/2406.07539

[46] A. Khazatsky, K. Pertsch, S. Nair, *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," 2024. [Online]. Available: https://arxiv.org/abs/2403.12945

[47] H. Walke, K. Black, A. Lee, *et al.*, "Bridgedata v2: A dataset for robot learning at scale," 2024. [Online]. Available: https://arxiv.org/abs/2308.12952

[48] A. Brohan, N. Brown, J. Carbajal, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," 2023. [Online]. Available: https://arxiv.org/abs/2212.06817

[49] T. Yu, D. Quillen, Z. He, R. Julian, A. Narayan, H. Shively, A. Bellathur, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," 2021. [Online]. Available: https://arxiv.org/abs/1910.10897

[50] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *arXiv preprint arXiv:2306.03310*, 2023.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[52] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[53] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015. [Online]. Available: https://arxiv.org/abs/1405.0312

[54] G. Aydemir, W. Xie, and F. Güney, "Self-supervised object-centric learning for videos," 2023. [Online]. Available: https://arxiv.org/abs/2310.06907

[55] K. Fan, Z. Bai, T. Xiao, T. He, M. Horn, Y. Fu, F. Locatello, and Z. Zhang, "Adaptive slot attention: Object discovery with dynamic slot number," 2024. [Online]. Available: https://arxiv.org/abs/2406.09196

[56] A. Didolkar, A. Zadaianchuk, R. Awal, M. Seitzer, E. Gavves, and A. Agrawal, "Ctrl-o: Language-controllable object-centric visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 29 523–29 533.