

A Related Work

Non-rigid object manipulation. Recent advances tackle individual tasks requiring more complex manipulation than traditional pick-and-place-style tasks, such as cutting [19, 20, 23, 24], peeling [25–27], and stir-frying [28]. However, these efforts tackle each task in *isolation*, often focus on the task’s mechanical aspects, lack general-purpose vision feedback, or rely heavily on simulation. In contrast, our work targets a broad class of spatially transformative tasks that require reasoning over visual state changes rather than contact dynamics alone, exploiting a *unified* visual representation that is shared across objects and state-change tasks.

Visual representations for robot learning. To accelerate downstream policy learning, recent works pretrain visual representations on large-scale data [12, 29–31]. More relevant to our novel visual rewards, VIP [31] learns an implicit value function over egocentric videos, while its extension LIV [12] further incorporates language-goal embeddings. There is also growing interest in LLMs [32] & VLMs [33] for robotic reasoning, typically using frame-level goal matching or symbolic planning. In contrast, SPARTA leverages a VLM for spatial reasoning over localized object regions, enabling dense reward generation and supporting both efficient planning and online RL for visually complex manipulation.

Affordances in robotics. Understanding *how* and *where* to interact with objects has driven a surge of interest in affordance-based functional grasping [34–39]. Parallel efforts in computer vision predict hand-object interactions [40–42], but they emphasize pick-and-place or grasping tasks. In contrast, we tackle a fundamentally different class of affordance—spatially evolving, visual object state transformations that generalize across tasks and robot embodiments. To our knowledge, this represents the first affordance reasoning approach for such manipulations achieving non-rigid object interactions on a real robot.

Object state change understanding. OSC is explored in computer vision for video-level classification [6, 7], segmentation [8, 43, 44], and generation [45]. Our work is inspired in part by the *spatially progressing object state change* (SPOC) task [8], which segments state-changing objects into actionable and transformed regions. Trained on large-scale instructional “how-to” videos [46], SPOC exhibits robust spatial reasoning across diverse objects and transformations. However, these models are vision-only: they passively analyze state changes but do not inform robot control. Our work bridges that gap. By integrating vision-based OSC understanding into robot manipulation, we show how robots can learn to act using SPOC-style affordances capturing gradual visual progress—difficult to address with tactile sensing [27], force models, or binary state classifiers [2, 47].

Real-world Reinforcement Learning Real-world RL enables autonomous policy learning directly from real-world interactions, avoiding the need for explicit world models or high-fidelity simulators. This makes it particularly promising for contact-rich manipulations, where accurate modeling is notoriously difficult [48, 49]. However, when tasks require progressive object state changes, existing methods struggle on two fronts: first, learning visual representations that capture subtle intra-object changes; and second, defining reward functions that provide dense, informative feedback [11, 16]. These challenges lead to poor sample efficiency and hinder real-world applicability. Our work tackles both issues by leveraging spatially progressing OSC segmentation maps, leading to successful policy learning on challenging tasks.

B Implementation Details

Tool-use primitives. At each predicted 2D location, high-level actions are executed via simple task-specific primitives following prior work [50, 51]: in-plane brush strokes with periodic refills for *spreading*, lateral motion plus downward press for *mashing* and *slicing*, with tools lifted after each action to prevent occlusion.

Robot platform. All experiments are conducted on a Franka Emika Panda robot, a 7-DoF collaborative manipulator equipped with torque sensing in each joint and a 3-finger parallel gripper. Its

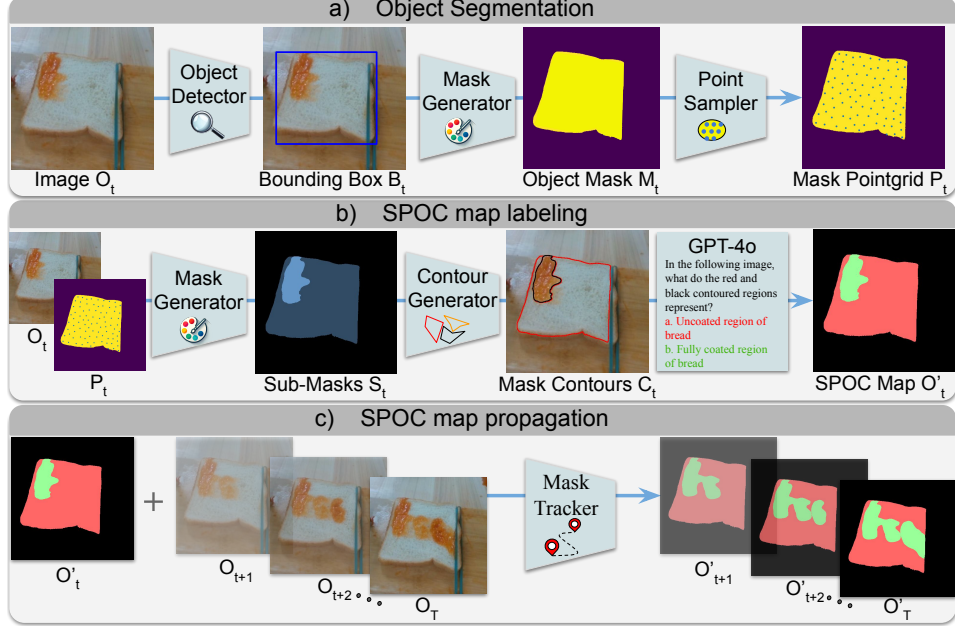


Figure 7: **Our SPOC affordance map generation pipeline.** (a) Grounded-SAM [52] is used to extract an object mask from the initial frame. (b) Farthest-point sampling generates intra-object regions, classified into *actionable* or *transformed* by prompting GPT-4o [14] using color-coded overlays. (c) Once classified, transformed regions are tracked across subsequent frames using DeAOT [15] to maintain temporal consistency with minimal computation.

precise joint control and compliant torque feedback make it well-suited for fine manipulation tasks such as spreading, mashing, and slicing. A front-facing camera provides RGB observations for the vision model.

Training. To keep training grounded in real-world constraints, we set episode lengths to match the natural granularity of each task. For spreading, episodes last 10 steps to reflect the smaller coverage per action, while for mashing and slicing, 5 steps suffice due to the broader area transformed by each action. All episodes begin from a fixed corner of the workspace for consistency. For SPARTA-L, we train policies with short real-world budgets: spreading is trained for 40 episodes (~ 3 hours at 1 Hz, including brush refills and resets), while mashing and slicing converge within ~ 1.5 hours thanks to shorter episodes. To simplify resets, we use clay proxies for mashing and slicing, and bootstrap exploration with a handful (~ 5) greedy rollouts, which stabilize early training. For policy learning, we adopt asynchronous SAC from SERL [16], finding that an actor-to-critic update ratio of 1:10 yields the best balance between policy improvement and stable value estimation. Other hyperparameters follow standard practice (learning rate $3e-4$ with warmup, $\gamma = 0.95$, reward weights $\alpha = 1, \beta = 1, \eta = 0.001$). Visual inputs are encoded via a ResNet-10 backbone, and proprioceptive inputs through a two-layer MLP. The same training protocol is applied across our method and baselines to ensure fair comparison.

C Integrating SPOC for Robotics

We adapt SPOC for robotics by generating SPOC affordance maps directly from real-time visual observations. While prior work [8] leverages Grounded-SAM [52] and CLIP [53], we find that replacing CLIP with a stronger vision-language model (VLM) such as GPT-4o [14] significantly improves segmentation accuracy—particularly in distinguishing intra-object regions (e.g., partially mashed banana). Instead of assigning a single label to the entire object mask, we sample multiple intra-object regions using farthest-point prompts with SAM, and classify each via GPT-4o into actionable or transformed states. Since per-frame GPT queries are slow (~ 5 s), we introduce a fast mask propagation strategy using DeAOT [15] tracking (~ 0.2 s/frame) to boost real-time throughput

for robotic control. See Fig. 7 for the full pipeline. These affordance maps offer dense, object-centric structure that is crucial for shaping progress-based rewards and guiding spatial-aware policy learning.

D Limitations and Future Work

While effective, SPARTA also reveals open challenges that suggest avenues for future research. First, our approach depends on SPOC affordance maps, which can occasionally exhibit noise or tracking inconsistencies—especially during fine-grained transitions. Nonetheless, we do observe some policy robustness to those errors, due to repeated exposure to accurate predictions and the dense reward formulation, which allows learning to proceed even when intermediate frames are noisy, as long as progress is eventually captured. Future work can explore vision segmentation model enhancements. Second, although policies generalize to new objects, performance degrades on unseen geometries—for example, a policy trained on rectangular slices may struggle with circular tortillas. Addressing this gap calls for shape-aware training or augmented experience. Third, our current pipeline avoids occlusion by only capturing visual inputs when the end-effector lifts between actions, precluding continuous perception during contact. Developing occlusion-resilient, contact-aware visual reasoning remains an open challenge. Overall, SPARTA demonstrates that progress-aware affordances can unlock a family of object state manipulations essential for everyday tasks, charting a path beyond rigid-body control.